Le poids des mots

Volume I

Le poids des mots

Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles

Proceedings of the 7th International Conference on Textual Data Statistical Analysis

Louvain-la-Neuve
10-12 mars 2004 / March 10-12, 2004

Gérald Purnelle, Cédrick Fairon, Anne Dister (éds)

Volume I

UCL PRESSES UNIVERSITAIRES DE LOUVAIN

Comité de programme / Program Committee

Ramon Alvarez	Universitad de Leon, SP
Harald Baayen	Université de Nimègue, NL
Simona Balbi	Università di Napoli, IT
Valérie Beaudouin	France Télécom R&D, FR
Monica Bécue	Universitat Politècnica de Catalunya, SP
Vincent Blondel	Université catholique de Louvain, BE
Sergio Bolasco	Università degli studi di Roma La Sapienza, IT
Étienne Brunet	Université de Nice Sophia Antipolis, FR
Joseph Denooz	Université de Liège, BE
Jean-Claude Deroubaix	Université libre de Bruxelles, BE
Anne Dister	Université catholique de Louvain, BE
Michel Dorban	Université catholique de Louvain, BE
Annibale Elia	Università di Salerno, IT
Étienne Évrard	Université de Liège, BE
Cédrick Fairon	Université catholique de Louvain, BE
Serge Fleury	Université Paris 3 Sorbonne Nouvelle, FR
Corinne Gobin	FNRS, Université libre de Bruxelles, BE
Sylvianne Granger	Université catholique de Louvain, BE
Michel Kerboal	INSERM, Université de Rennes 1, FR
Dominique Labbé	Université de Grenoble, FR
Ludovic Lebart	CNRS, ENST Paris, FR (Président)
Alain Lelu	Université de Franche-Comté, FR
Sylvie Mellet	CNRS, Nice, FR
Piet Mertens	Katholieke Universiteit Leuven, BE
Annie Morin	IRISA, Université de Rennes 1, FR
Gérald Purnelle	Université de Liège, BE
Martin Rajman	EPFL Lausanne, CH
Max Reinert	CNRS, Université de Versailles SQY, FR
Marco Saerens	Université catholique de Louvain, BE
André Salem	Université Paris 3 Sorbonne Nouvelle, FR
Pascale Sebillot	IRISA, Université de Rennes 1, FR
Fiona Tweedie	University of Glasgow, RU
François Yvon	ENST, FR

Comité d'organisation / Organization Committee

Anne Dister	CENTAL, Université catholique de Louvain, BE
Cédrick Fairon	CENTAL, Université catholique de Louvain, BE
Gérald Purnelle	CIPL-LASLA, Université de Liège, BE

http://cental.fltr.ucl.ac.be http://www.ulg.ac.be/cipl/

Organisation locale / Local Organization

Bernadette Dehottay, Claude Devis, Michel Thomas, Laurent Simon, Patrick Watrin.
Conception du CD-Rom JADT 2004 : Parick Watrin.

Graphisme de la couverture et des publications JADT 2004 : Olivier Vereecken http://aphine.com

Dépôt légal : D/2004/9964/10
ISBN : 2-930344-49-0

Imprimé en Belgique

Diffusion :
www.i6doc.com, l'édition universitaire en ligne

Sur commande en librairie ou à :
Diffusion universitaire CIACO
Grand-Place, 7
1348 Louvain-la-Neuve, Belgique

Tél. 32 10 47 33 78
Fax 32 10 45 73 50
duc@ciaco.com

Table des matières / Table of Contents

Volume I

Volume II

Conversation text types: A multi-dimensional analysis

Douglas Biber

Northern Arizona University, USA
douglas.biber@nau.edu

Abstract

Multi-dimensional (MD) analysis is a methodological approach that applies multivariate statistical techniques (especially factor analysis and cluster analysis) to the investigation of register variation in a language. The approach was originally developed to analyze the full range of spoken and written registers in a language. Early studies focused on English register variation (Biber 1985, 1986 and 1988), while later studies have applied the same approach to Somali, Korean, Tuvaluan, Taiwanese, and Spanish.

Surprisingly, these studies have found some striking similarities in the underlying 'dimensions' that distinguish among spoken and written registers in these diverse languages. It is even more surprising that MD studies of restricted discourse domains have also uncovered dimensions that are similar in linguistic form and function to the more general studies of register variation.

The present study presents an MD analysis of a single register: conversation. Three primary dimensions of variation are identified, and then cluster analysis is used to distinguish among six conversation text types. The dimensions and text types are interpreted in linguistic and functional terms.

The author's expectations were that a unique set of dimensions would emerge to characterize the variation among conversational texts. Instead, the three dimensions identified here turn out to be closely related to dimensions identified in previous analyses of general register variation. Taken together with previous studies, the present study of conversation raises the possibility of universal dimensions of variation.

1. Introduction

Multi-dimensional (MD) analysis is a methodological approach that applies multivariate statistical techniques (especially factor analysis and cluster analysis) to the investigation of register variation in a language. The approach was originally developed to analyze the range of spoken and written registers in English (Biber 1985, 1986 and 1988). There are two major quantitative steps in an MD analysis: (1) identifying the salient linguistic co-occurrence patterns in a language; and (2) comparing spoken and written registers in the linguistic space defined by those co-occurrence patterns. In a third step, it is possible to identify groupings of texts — 'text types' — that are maximally similar in their multi-dimensional profiles.

Almost any linguistic feature will vary in its distribution across registers, reflecting the discourse functions of the feature in relation to the situational characteristics of each register (see, e.g., the grammatical descriptions in the *Longman Grammar of Spoken and Written English*; Biber *et al.*, 1999). However, individual features cannot reliably distinguish among registers: There are simply too many different linguistic characteristics to consider, and individual features often have idiosyncratic distributions. Instead, analyses based on linguistic *co-occurrence* and *alternation* patterns are required to uncover general register differences.

The theoretical importance of linguistic co-occurrence has been emphasized by linguists such as Firth, Halliday, Ervin-Tripp, and Hymes. Brown and Fraser (1979: 38-39) observe that it

can be 'misleading to concentrate on specific, isolated [linguistic] markers without taking into account systematic variations which involve the co-occurrence of sets of markers'. Ervin-Tripp (1972) and Hymes (1974) identify 'speech styles' as varieties that are defined by a shared set of co-occurring linguistic features. Halliday (1988: 162) defines a register as 'a cluster of associated features having a greater-than-random...tendency to co-occur'.

The MD approach gives formal status to the notion of linguistic co-occurrence, by providing empirical methods to identify and interpret co-occurrence patterns as underlying *dimensions* of variation. The co-occurrence patterns comprising each dimension are identified quantitatively through factor analysis. It is not the case, though, that quantitative techniques are sufficient in themselves for MD analyses of register variation. Rather, qualitative techniques are required to interpret the functional bases underlying each set of co-occurring linguistic features. The dimensions of variation have both linguistic and functional content. The linguistic content of a dimension comprises a group of linguistic features (e.g., nominalizations, prepositional phrases, attributive adjectives) that co-occur with a high frequency in texts. Based on the assumption that co-occurrence reflects shared function, these co-occurrence patterns are interpreted in terms of the situational, social, and cognitive functions most widely shared by the linguistic features. That is, linguistic features co-occur in texts because they reflect shared functions.

Several experiments have been carried out to evaluate the reliability (and to a lesser extent validity) of the original MD analysis of register variation in English. For example, Biber (1990) shows that factor analyses carried out on split corpora result in nearly the same dimensions of variation, as long as the texts in those corpora are sampled to include equivalent ranges of register variation. Biber (1993) shows how these dimensions can be used to predict the register category of unclassified texts with a high degree of accuracy (using discriminant analysis). And Biber (1992) uses confirmatory factor analysis to test the goodness of fit of several factorial models determined on theoretical grounds, confirming the basic structure identified using exploratory factor analysis in the 1988 analysis.

While early MD studies focused on register variation in English, subsequent studies have applied the same approach to Somali, Korean, Tuvaluan, Taiwanese, and Spanish (see, e.g., Biber, 1995; Jang, 1998). Although these studies all apply the same methodological approach, they are carried out independently. In each case, a corpus was designed to represent the range of spoken and written registers found in the target culture, and a computational tagger was written to capture the grammatical structure of the target language. The set of linguistic variables used in each analysis includes the full range of lexical/grammatical distinctions that are relevant in the target language. Despite this fact, the resulting MD analyses have turned out to be strikingly similar in some respects. In particular, the analyses of all languages have uncovered dimensions relating to interactiveness/involvement versus informational focus, the expression of personal stance, and narrative versus non-narrative discourse (see Biber, 1995, especially Chapter 7).

The MD methodological framework has also been applied to more restricted discourse domains.[1] These include analyses of elementary school registers (Reppen, 1994 and 2001),

[1] There have also been several studies of specific registers that apply the dimensions that were identified and interpreted in the 1988 MD analysis of spoken and written variation in English (see, for example, the collection of studies in Conrad and Biber, 2001). It is important to note that these studies do *not* entail separate MD analyses. That is, these studies apply the dimensions identified in the 1988 MD analysis of English to some new discourse domain, but they do not undertake new MD analyses (i.e., involving a new factor analysis).

job interview language (White, 1994), television talk shows (Connor-Linton, 1989), 18th century written and speech-based registers (Biber, 2001), university spoken and written registers (Biber, 2003), and academic subregisters (e.g., Grabe, 1987; Kanoksilapatham, 2003). Many of these studies have identified dimensions of variation similar to those found in the cross-linguistic studies, especially relating to the same functional concerns of interactiveness/involvement versus informational focus, the expression of personal stance, and narrative versus non-narrative discourse.

This result is surprising for two reasons. First, the statistical technique of factor analysis — like all correlational techniques — requires variability. Two linguistic variables cannot be shown to correlate unless the texts included in an analysis represent a wide range of variation for those variables. Similarly, factor analysis cannot reliably identify sets of co-varying linguistic features unless the texts included in the analysis represent a wide range of variation for the full set of features. Thus, factor analysis is most appropriate for general analyses of spoken and written texts, which represent an extensive range of variation for almost any linguistic feature (see the detailed analyses in Biber *et al.*, 1999). In contrast, it might be assumed that factor analysis is less appropriate for analyses of texts from a single, restricted discourse domain, because that domain will represent a much smaller range of variation.

Second, to the extent that there is linguistic variability among the texts in a restricted discourse domain, there is no reason to assume that it would be similar to the patterns of variation found in a general-purpose corpus. We would rather expect to find different linguistic features varying in a restricted domain, reflecting the specific functional differences found in that domain. In the MD analyses, these specific patterns of linguistic variation should result in dimensions of variation that are unique to each discourse domain.

Previous MD analyses have shown that restricted discourse domains represent sufficient linguistic variability for the successful application of this methodological approach. More surprisingly, these analyses show that some of the same basic dimensions of variation seem to be fundamentally important across restricted and general discourse domains. (In addition, there are other dimensions that are unique to a particular domain.) This repeated finding — that some dimensions occur across languages and across general and restricted discourse domains — raises the possibility of universal dimensions of register variation.

The present study further explores this possibility by undertaking an MD analysis of linguistic variation within a single spoken register: conversation. Factor analysis is used to identify the linguistic dimensions of variation operating in this discourse domain, and then cluster analysis is used to identify conversation 'text types' that are well-defined in that multi-dimensional space. The following sections describe these analyses, followed by discussion of the more general theoretical implications for the study of register variation.

2. Overview of methodology in the Multi-Dimensional approach

A Multi-Dimensional analysis follows eight methodological steps:

1. An appropriate corpus is designed based on previous research and analysis. Texts are collected, transcribed (in the case of spoken texts), and input into the computer. The situational characteristics of each spoken and written register are noted (e.g., communicative purpose, production circumstances, etc.).

2. Research is conducted to identify the linguistic features to be included in the analysis, together with functional associations of the features.

3. Computer programs are developed for automated grammatical analysis, to identify — or 'tag' — all relevant linguistic features in texts.

4. The entire corpus of texts is tagged automatically by computer, and all texts are edited interactively to insure that the linguistic features are accurately identified.

5. Additional computer programs compute normed counts of each linguistic feature in each text of the corpus.

6. The co-occurrence patterns among linguistic features are analyzed, using factor analysis.

7. The factors are interpreted functionally as underlying dimensions of variation.

8. Dimension scores for each text are computed; the mean dimension scores for each register are then compared to analyze the salient linguistic similarities and differences among the registers being studied. The functional interpretation of each dimension is refined based on the distribution among registers.

3. Preliminary steps for the MD analysis of conversation: Corpus and linguistic features

3.1. *The conversation corpus*

The corpus used for the present analysis is taken from the Longman Spoken and Written English Corpus (LSWE Corpus; see Biber *et al.*, 1999: chapter 1). Only the British English sub-corpus of conversation was analyzed here; this sub-corpus includes 164 texts containing c. 4 million words. (A large part of this corpus was also included in the BNC sample of conversation.) Texts were collected by asking participants to carry tape recorders for several days, recording their daily interactions. The language collected in this way is conversational, but most text files are very large and actually include many different conversations. Participants would generally turn the tape recorder off in between conversations, but each text file in the corpus includes all the conversations that were recorded on a single tape.

The LSWE/BNC corpus of conversation is large enough to provide the basis for a multi-dimensional analysis. Texts were collected over many days by many different participants, representing a wide range of social backgrounds. As a result, the corpus should represent the range of linguistic variability found within conversation. However, much of that variability is lost when the corpus is analyzed in its current form, with each text file combining multiple conversations. Individual conversations can vary with respect to situation and purpose, and to the extent that there is linguistic variability among conversational texts, it will be associated with those situational/communicative differences. As a result, the first step in the present analysis was to segment text files into individual conversations (based on the internal headers included in each text file). Table 1 shows that the 164 text files in the LSWE conversational corpus were segmented into 2,926 individual conversations. 760 of these conversations were shorter than 200 words. Because of the difficulties in obtaining reliable rates of occurrence for linguistic features in shorter texts, these shorter conversations were excluded from subsequent analysis.

Table 1 shows that the conversations included in the analysis are on average quite long (1,775 words), with the longest conversations being almost 14,000 words.

# of text files: 164 (conversation transcripts from the LSWE Corpus) # of words: 3,930,000
Individual conversations longer than 200 words: 2,166 Individual conversations shorter than 200 words: 760 (dropped from subsequent analyses) Total individual conversations: 2,926
Length of individual conversations included in the analysis: mean = 1,775 words min = 200 words max = 13,776 words

Table 1. Initial segmentation of the conversation corpus into individual conversations

3.2. *Linguistic features used for the analysis*

After the conversation corpus was segmented, each conversation was automatically 'tagged' using the Biber grammatical tagger. The current version of this tagger incorporates the corpus-based research carried out for the *Longman Grammar of Spoken and Written English* (Biber *et al.*, 1999). The tagger identifies a wide range of grammatical features, including word classes (e.g., nouns, modal verbs, prepositions), syntactic constructions (e.g., WH relative clauses, conditional adverbial clauses, that-complement clauses controlled by nouns), semantic classes (e.g., activity verbs, likelihood adverbs), and lexico-grammatical classes (e.g., that-complement clauses controlled by mental verbs, to-complement clauses controlled by possibility adjectives). Appendix A lists the full set of features analyzed here.

4. Identifying and interpreting the dimensions of variation in English conversation

As noted above, the Multi-Dimensional approach to register variation uses factor analysis to reduce a large number of linguistic variables to a few basic parameters of linguistic variation. In MD analyses, the distribution of individual linguistic features is analyzed in a corpus of texts. Factor analysis is then used to identify the systematic co-occurrence patterns among those linguistic features — the 'dimensions' — and then texts and registers are compared along each dimension.

Table 2 gives the full factorial structure for the analysis in this case, while Table 3 summarizes the important linguistic features defining each dimension (i.e., features with factor loadings over + or –.3). Only 27 of the original 120+ linguistic features were retained in the final factor analysis. Several features were dropped because they were redundant or overlapped to a large extent with other features. For example, the counts for common verbs, nouns, and adjectives overlapped extensively with the semantic categories for those word classes, even though the counts were derived independently. In other cases, features were dropped because they were extremely rare in conversation. Several of these features were combined into a more general class. For example, the seven phrasal verb types were combined into a single feature. Similarly, the five specific types of postnominal modifying clause were combined into a single 'relative clause' feature. Finally, many features were dropped either because they did not vary across conversational texts, or because they shared little variance with the overall factorial structure of this analysis (as shown by the communality estimates). The solution for three factors was selected as optimal. These three factors account for only 36% of the shared variance, but they are readily interpretable, and subsequent factors accounted for relatively little additional variance. Given that only 27 linguistic variables were retained in the final factor analysis, the solution with 3 factors was considered optimal.

	Factor1	Factor2	Factor3
Major Factor 1 Features:			
wrdlngth	0.75638	0.06122	-0.08615
n_nom	0.52756	0.02817	-0.08707
prep	0.46987	0.04783	0.05038
abstrctn	0.46893	0.13866	-0.11843
rels	0.44690	0.16282	-0.09168
adj_attr	0.35287	-0.17778	-0.08123
allpasv	0.29913	0.08863	0.09965
contrac	-0.43589	0.24817	-0.18706
pro1	-0.39020	0.23735	0.08420
pro2	-0.36418	-0.07955	-0.25500
actv	-0.31900	-0.15923	0.03750
Major Factor 2 Features:			
that_del	-0.02783	0.67341	0.38349
mentalv	-0.01213	0.66432	-0.07223
fact_vth	0.14676	0.54415	0.08570
lkly_vth	0.01636	0.42516	0.01091
lklyadvl	0.38093	0.40397	-0.07723
sub_all	0.07570	0.35545	-0.00143
gen_hdg	0.33564	0.34281	-0.08784
factadvl	0.28542	0.33556	0.01889
n	0.16827	-0.52004	-0.04842
wh_ques	-0.23266	-0.34870	-0.10075
Major Factor 3 Features:			
pasttnse	-0.04272	-0.04189	0.79494
nonf_vth	-0.12568	0.14077	0.60910
commv	-0.13631	0.06964	0.58285
pro3	0.00932	0.10243	0.52077
pres	-0.45977	0.26021	-0.51128
allmodal	-0.23413	0.20474	-0.26521
Inter-Factor Correlations			
	Factor1	Factor2	Factor3
Factor1	1.00000	-0.24213	0.09272
Factor2	-0.24213	1.00000	0.08185
Factor3	0.09272	0.08185	1.00000

Table 2. Results of the factor analysis: 3 factor solution; Promax rotation.

Each factor comprises a set of linguistic features that tend to co-occur in the conversations from the conversation corpus. Factors are interpreted as underlying 'dimensions' of variation based on the assumption that linguistic co-occurrence patterns reflect underlying communicative functions. That is, particular sets of linguistic features co-occur frequently in texts because they serve related communicative functions. Features with positive and negative loadings represent two distinct co-occurrence sets. These comprise a single factor because the two sets tend to occur in complementary distribution: when a conversation has high frequency

of the positive set of features, that same conversation will tend to have low frequencies of the negative set of features, and vice versa. In the interpretation of a factor, it is important to consider the likely reasons for the complementary distribution between positive and negative feature sets as well as the reasons for the co-occurrence patterns within those sets.

Dimension 1: Information-focused vs. interactive discourse

Features with positive loadings: word length, nominalizations, prepositional phrases, abstract nouns, relative clauses, attributive adjectives, passive verb phrases, (likelihood adverbs, general hedges)

Features with negative loadings: present tense verbs, contractions, 1st person pronouns, 2nd person pronouns, activity verbs

Dimension 2: Stance vs. context-focused discourse

Features with positive loadings: *that*-deletions, mental verbs, factual/mental verb + *that*-clause, likelihood/mental verb + *that*-clause, likelihood adverbs, adverbial clauses, general hedges,

factual adverbs

Features with negative loadings: nouns, *WH*-questions

Dimension 3: Narrative-focused discourse

Features with positive loadings: past tense verbs, 3rd person pronouns, non-factual/communication verb + *that*-clause, communication verbs, *that*-deletions

Features with negative loadings: present tense verbs

Table 3. Summary of the factorial structure

For example, the positive features on Factor 1 (e.g., long words, nominalizations, prepositional phrases, abstract nouns, relative clauses, etc.) all relate to informational purposes. These features are mostly associated with elaborated noun phrases and a dense integration of information in a text; previous MD studies have shown these features to be typical of written non-fictional registers intended for specialist audiences (see, e.g., Biber, 1995; Biber and Finegan, 2001).

In contrast, the negative features on Dimension 1 reflect a focus on the immediate interaction and activities: present tense verbs, contractions, 1st and 2nd person pronouns, and activity verbs. The overall interpretation of Dimension 1 is thus relatively straightforward, showing that conversations tend to be either 'informational' or 'interactive', but not both. The functional label 'Information-focused versus interactive discourse' is proposed for this dimension.

The positive features on Dimension 2 are mostly linguistic features that express 'stance': personal attitudes or indications of likelihood. In the 1988 MD study of spoken and written register variation, several of these features were shown to co-occur typically with interactive and reduced structure features (on Dimension 1). In contrast, the analysis here shows that stance-focused discourse is not necessarily highly interactive discourse, and vice versa. (This dimension also includes several specific features that were not distinguished in the feature set used for the 1988 analysis, such as likelihood/mental verb + that-clause and factual adverbs).

The negative pole of Dimension 2 shows a surprising co-occurrence of only two features: nouns and WH-questions. In past analyses, nouns have co-occurred with other stereotypically 'literate' features (like adjectives, prepositional phrases, etc.), while WH-questions have co-occurred with stereotypically 'oral' and interactive features. The interpretation here must consider why these two features would tend to co-occur in conversations, and why they would

tend to occur in a complementary distribution to stance features. Consideration of texts with a high frequency of these two features indicates that they are used together to reflect a focus on the larger context. WH-questions — the 'what', who', 'where', 'when' and 'how' — directly ask about that context, and nouns are the primary device used to refer to it. Thus, considering both positive and negative poles, we propose the interpretive label 'Stance-focused versus context focused discourse' for Dimension 2.

Finally, Dimension 3 is composed of stereotypically narrative features — past tense verbs, 3rd person pronouns, and communication verbs controlling that-clauses; the only negative feature on this dimension is present tense verbs. Given this grouping of features, the interpretation as 'Narrative-focused discourse' is uncontroversial.

5. Identifying and interpreting conversation text types

Most MD studies have been undertaken to investigate the patterns of variation among 'registers': varieties of language that are defined by their situational (i.e. non-linguistic) characteristics (see Biber, 1994). Conversation is an example of a register according to this definition, as is newspaper reportage, classroom lectures, personal letters, and academic research articles. Registers can be defined at any level of specificity, depending on the extent to which the situational characteristics are specified. For example, academic prose is a very general register, while academic research articles, psychology research articles, and methodology sections in experimental psychology research articles are registers defined at increasing levels of specificity. The original MD studies (Biber, 1986 and 1988) analyzed a wide range of general spoken and written registers in English, while many subsequent analyses have applied those dimensions to the analysis of other more specialized registers (see, e.g., the studies in Conrad and Biber 2001).

These analyses have shown that there are important, systematic linguistic differences among registers. Those linguistic differences exist because of the functional basis of MD analysis: linguistic co-occurrence patterns reflect underlying communicative functions. Registers differ in their situational/communicative characteristics, and as a result, the dimensions identify important linguistic differences among registers. However, it is important to note that the register categories are defined in situational rather than linguistic terms.

A complementary perspective on textual variation is to identify and interpret the text categories that are **linguistically** well defined, referred to as **text types**. Text type distinctions have no necessary relation to register distinctions. Rather, text types are defined such that the texts within each type are maximally similar in their linguistic characteristics, regardless of their situational/register characteristics. However, because linguistic features have strong functional associations, text types can be interpreted in functional terms.

Text types and registers thus represent complementary ways to dissect the textual space of a language. Text types and registers are similar in that both can be described in linguistic and in situational/functional terms. However, the two constructs differ in their primary bases: registers are defined in terms of their situational characteristics, while text types are defined linguistically.

In the MD approach, text types are identified quantitatively using Cluster Analysis, with the dimensions of variation as predictors. Cluster analysis groups texts into 'clusters' on the basis of shared multi-dimensional/linguistic characteristics: the conversations grouped in a cluster are maximally similar linguistically, while the different clusters are maximally distinguished. This approach has been used to identify the general text types in English and Somali (see

Biber, 1989 and 1995). The present section describes the text types that can be distinguished linguistically within the single register of conversation.

The dimensions of variation (see Section 4 above) are used as linguistic predictors for the clustering of conversations. The individual feature counts are first standardized so that each feature has a comparable scale with a mean of 0.0 and a standard deviation of 1. (The standardization was based on the overall means and standard deviations for each feature in the conversation corpus.) Then, 'dimension scores' were computed by summing the standardized frequencies for the features comprising each of the three dimensions. The cluster analysis is based on the three dimension scores for each conversation.

The methodology in this analytical step can be illustrated conceptually by the 2-dimensional plot in Figure 1. Each point on Figure 1 represents a conversation, plotting the scores for that conversation on Dimensions 1 and 2. The numbers in the figure show the cluster number for each conversation, based on the results of the cluster analysis. Conversations that are similar in their dimension scores are grouped together as a cluster, or 'text type'. For example, the conversations labelled with a '1' on Figure 1 all have large positive scores on Dimension 1 (the vertical axis) and large negative scores on Dimension 2 (the horizontal axis). In contrast, Cluster 2 has positive scores on both Dimensions 1 and 2.

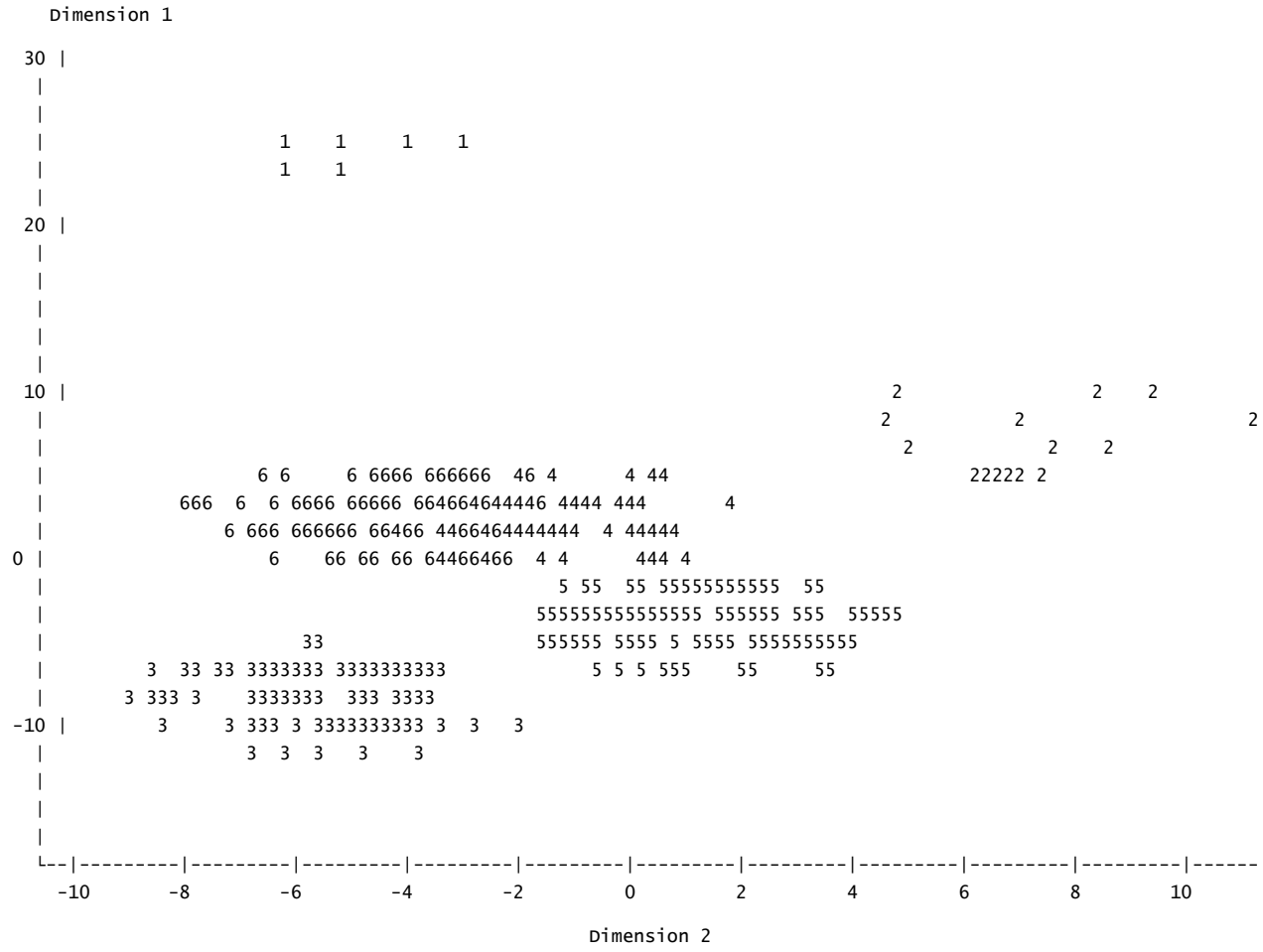


Figure 1. Plot of VBDUs along Dimension 1 vs. Dimension 2 (showing all DUs with a distance < 3 from the cluster centroid. Symbol is value of CLUSTER; NOTE: 194 obs hidden.)

Cluster analysis performs this grouping statistically, based on the scores for all three dimensions. Figure 1 shows the distribution across only two dimensions (1 and 3); these two dimensions were chosen because they provide a good visual display of how the conversations within each cluster are grouped based on their dimension scores. However, the actual cluster analysis uses all three dimension scores to identify the groupings of conversations that are maximally similar in their linguistic characteristics.

Cluster analysis is an exploratory statistical technique. The FASTCLUS procedure from SAS was used for the present analysis. Disjoint clusters were analyzed because there was no theoretical reason to expect a hierarchical structure. Peaks in the Cubic Clustering Criterion and the Pseudo-F Statistic (produced by FASTCLUS) were used to determine the number of clusters. These measures are heuristic devices that reflect goodness-of-fit: the extent to which the texts within a cluster are similar, while the clusters are maximally distinguished. In the present case, these measures had peaks for the 3-cluster solution and for the 6-cluster solution. The latter was chosen for subsequent analyses because it provided greater discrimination among the specialized clusters, facilitating the interpretation of those clusters as conversation text types.

Figure 1 shows the distribution of these six clusters in only a 2-dimensional space, whereas the cluster analysis is actually based on a 3-dimensional space. It turns out that the third dimension is also important in defining some clusters. For example, Cluster 4 is not sharply delimited in terms Dimensions 1 and 2, but all conversations in this cluster have large positive scores on Dimension 3 (‘narrative’).

Tables 4 and 5 provide a descriptive summary of the cluster analysis results. Table 4 shows the number of conversations grouped into each cluster, while Table 5 gives descriptive statistics for each dimension across the clusters. The clusters differ notably in their distinctiveness: the smaller clusters are more specialized and more sharply distinguished linguistically. For example, Cluster 1 has only 40 conversations; linguistically, the conversations grouped in Cluster 1 have extremely large positive scores on Dimension 1 (‘informational’); large negative scores on Dimension 2 (‘context-focused’); and scores near 0.0 on Dimension 3 (‘narrative’). At the other extreme, Cluster 5 is a ‘general’ text type: it is large (680 conversations) and relatively unmarked in its dimension scores.

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Nearest Cluster	Distance Between Cluster Centroids
1	40	4.6276	19.8319	2	18.2629
2	116	4.3710	16.9770	4	9.5460
3	496	3.2839	18.7622	5	8.6697
4	308	3.4828	18.2692	6	8.2551
5	680	3.2643	16.2853	4	8.3268
6	526	3.2447	17.4048	4	8.2551

Table 4. Summary of the Cluster Analysis

Cluster Means

Cluster	Dim. 1	Dim. 2	Dim. 3	
1	22.15	-5.08	-0.99	(Informational context-focused)
2	7.67	5.87	0.93	(Informational stance-focused)
3	-8.04	-5.19	-2.88	(Interactive context-focused)
4	2.12	-0.31	5.61	(Narrative)
5	-4.15	1.74	0.55	(Unmarked interactive)
6	2.63	-4.46	-1.50	(Unmarked context-focused)

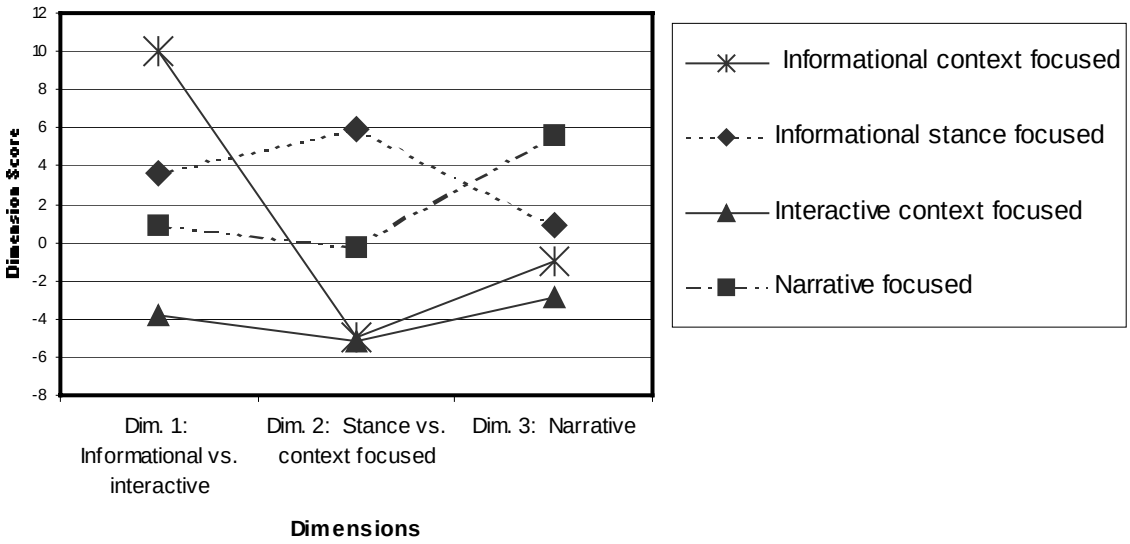
Cluster Standard Deviations

Cluster	Dim. 1	Dim. 2	Dim. 3
1	5.02	5.19	3.46
2	5.05	4.47	3.41
3	3.75	3.42	2.54
4	3.72	3.28	3.42
5	3.11	3.49	3.17
6	3.75	3.54	2.22

Table 5. Cluster descriptive statistics for each dimension

The clusters can be interpreted as Conversation Text Types, because each cluster represents a grouping of conversations with similar linguistic profiles. Figure 2 compares the linguistic characteristics of the four most distinctive of these conversation types, plotting their mean dimension scores. The 'general' conversation types — clusters 5 and 6 — are not plotted in Figure 2.

Figure 2: Multi-Dimensional profile for Conversational Text Types 1-4
(Note: Dimension 1 has been transformed to a scale of 10 for comparison)



Taken together, Table 4 and Figure 2 provide the basis for the interpretation of each conversation type. (These interpretations are refined by consideration of individual conversations from each type.)

Type 1 is the most specialized, with the fewest number of texts (only 40, or about 2% of the conversations in the corpus). Linguistically, these conversations are extremely informational (Dimension 1) and focused on the context (Dimension 2). Text Sample 1 provides an example of a conversation from this cluster. This text illustrates the dense use of 'informational' features, such as nominalizations (e.g., *conversation, sophistication, agreement, possibility, information*), other long words (e.g., *paperwork, computer-wise, computerized, consequently*), attributive adjectives (*modern, massive, great, preliminary, certified*), passives (*be/getting inundated*), prepositional phrases (*to you, about the conversation, with Alec, on a piece of paper*), and relative clauses (*things that you're liable to get asked*). Although texts from this cluster would be considered interactive and involved in comparison to written expository texts, they are highly informational in comparison to other conversational texts.

Text Sample 1 Conversation from Cluster 1: Informational, Context-focused

A: I, I want to talk to you about er the conversation I had with Alec <name> yesterday, he seems to be inundated with having to get details about <unclear> on his er, all his paperwork and so on, and he seems to be inundated and he sounded a bit low, quite frankly, to me yesterday on the phone that he was getting inundated with all this
B: Mm, mm
A: work. I said I'm quite sure there must be something that could be done computer-wise
B: Right
A: but he sort of pooh-poohed it and sort of said well you know, we're getting a bit too old for all this modern sophistication of computers and so on, well I said well quite frankly I am not totally in agreement with you, because as you probably know Clyde <name> was looking into a program which will could alleviate a lot
B: Yes I know, I know
A: of the work, that I do, but I
B: yes it's on the <unclear>
A: would tell you right here and now, er I'm still retaining my bible you know the book
B: Yeah, yes, yes
A: that I have downstairs, because it's, if it was to be computerized, it would be a massive great bloody great volume
B: Yes
A: and I would be carrying this around and it just wouldn't be feasible
B: Quite, right
A: so he said that apparently whenever he came back to B S H he was told by Neville roughly about eighteen hundred acres would be sort of his target
B: Target, right
A: and it's, it's multiplied by about three or four times that you see
B: Oh right, right, right
A: so consequently he's getting inundated, he really is apparently under pressure
B: Mm, mm, right
A: so this is why I raised the very conversation about it
B: Right, right
A: and er, I said well look I'll have a word with the erm, with <name> and see if he can think of anything that might
B: Yes alleviate the point
A: all things in mind that are possible on er, on er computer, and he said that he hadn't much time to think about it and said well look, maybe over Easter
B: Mm
A: put down on a piece of paper what essentials you want done
B: Right
A: and what things that you're liable to get asked
B: Right, mm, mm
A: so he's going to do that, so I said well look, do you mind if I had a wee sort of preliminary talk with him

B: Right
A: see if, if it's a possibility
B: Right
A: What he's looking for is certified numbers, field numbers
B: That sort of
A: all this sort of information

Type 2 is also relatively specialized (with only 116 conversations grouped into this cluster). Linguistically, this conversation type is relatively informational (Dimension 1) but especially marked for being highly stance-focused (Dimension 2). (This conversation type should be contrasted with Type 5: a much larger cluster that is stance-focused and highly interactive rather than informational.) Text Sample 2 illustrates the typical linguistic characteristics of Conversation Type 2. Notice especially the frequent mental verbs (e.g., *know, think, expect, want*), stance verbs controlling *that*-clauses, usually with the *that* omitted (e.g., *would have thought…, I think…, I suppose…*), and the frequent hedges and stance adverbs and adverbials (*surely, obviously, really, actually, probably, certainly, to be perfectly frank*).[2] Texts from this cluster are informational, in that they are focused on discussion of a particular topic rather than the immediate interpersonal interaction, but their primary purpose is the expression of personal stance in relation to that topic.

Text Sample 2. Conversation from Cluster 2: Informational, Stance-focused

A: No no no one person that's not right.
B: Oh, right.
A: There is no, statutory obligation for the person organizing it
C: Oh, I know.
B: Well not the organizer surely oh I know I would have thought you'd have to, <unclear> shoot it
A: I'm sure that the social services require psychiatric or
B: Mm, I would of thought so
A: obviously medical <unclear> what you're doing. Mhm but they're to be qualified people involved. But I would have expected that the whole thing would have to be operated by, somebody who was qualified.
B: I don't know, because like, you know like the doctors <unclear>
A: <unclear> I think it sort of depends how big that you want to get involved in. If you're just somebody who's on the outside providing services, to keep the smooth running of it then you don't really have to know anything about it.
C: Mm.
A: But if you're actually involved in it, and you want to be involved in the people, then I think you have to know something about it.
B: Well the other evening they were showing something on TV, one of these doctors', doctors' practices that are opting out or whatever. And they got a stockbroker, someone who used to be a stockbroker, actually managing the whole practice.
A: Yeah.
B: I mean he's obviously not qualified as a doctor.
A: Mhm.
B: So I mean I suppose they'll look at it in the same kind of way, somebody who's got managerial, management qualities rather than . — I suppose people who are interested in the other side of it, the medical side of it, probably, really be geared up to organizing the money side of it wouldn't they, usually one or the other.
C: So have you done any more calculations on it?
A: There's nothing really more <unclear> I mean the whole thing is a budget guesstimate. I've no idea yet, really what, I mean, you know, for instance I don't know how much ratio staff to patients they need, therefore you can't really, you know, follow that up because you've no idea what the costs themselves could be.
D: Well you don't know, have you, have you found the statutory requirement for space yet? Per person. —

[2] Note also the dense use of discourse markers (e.g., *I mean, you know*) supporting the expression of stance in this conversation.

A: I think the thing is going to come unstuck . — in the, I think the biggest thing is, I was thinking, is the fact that you've got to get <unclear> I wouldn't get a commitment from Social Services until they see a property actually ready for occupation. Now I'm not gonna be prepared to go through the whole business and then find them say oh sorry you're wrong.
C: Property is the biggest bugbear.
A: Yeah. Because I don't think
C: If you're actually sitting on <unclear>
A: I don't think the banks are gonna want to invest. To be perfectly frank. . — You see the only way we can get equity out and put money in ourselves is by selling this place.
C: Yes.
A: Therefore if we don't actually want to live in the same place as the residents, which I certainly wouldn't want to do, right. We'd have to buy two <unclear> adjoining -

In contrast, Type 3 conversations are much more common (496 conversations grouped into this cluster, or 23% of all conversations in the LSWE Corpus). These conversations are extremely interactive and focused on the immediate context, as illustrated by Text Sample 3. The turns in this conversation are short and highly interactive (notice the dense use of *I* and *you*), and there is a dense use of common nouns together with WH-questions to express context-dependent information.

Text Sample 3. Conversation from Cluster 3: Interactive, Context-focused

A: I'm coming home at lunchtime. There's milk on the step. Bye-bye.
B: But ... lunchtime
A: Right. We'll have to get cracking.
B: what d'ya mean lunchtime?
A: Well lunchtime I'm going to go in and pop back to get things. I've locked it, it was unlocked. Right.
C: Can I go in the front?
A: Tie your belt up please. Tie your belt up. — Okay, speedily . — now
B: Oh crash, bang, wallop you're a
A: but both doors were open, you know. Start the car.
C: <screaming>Ah! You happy
A: See the co=
C: now?
A: can you er zip your zips up please? Keira. Can you zip your zip up?
B: I can't.
A: What do you think you'll be doing at school today?
B: Recorder concert!
A: Oh! Have you got your recorder? In school?
B: No! Er, yes, yes
A: Yeah.
B: yes.
A: Now, what you gonna be playing?
B: Joe Joe stubbed his toe. Joe Joe stubbed his toe and ... Indian Warrior.
A: Oh!
B: <singing>Big chief, Indian warrior, warrior, warrior. Big chief, Indian warrior. High ... ho! High ... ho! High ... ho! oh! oh!
A: Right.
B: And erm ... the skateboard ride.
A: <crunches gears> Ooh! That gear. Keeps changing with the
B: Mummy. You know what I've
A: Skateboard ride?
B: you know what, that I
A: What's that one?
B: ca= just can play that, I couldn't do recorders that well?
A: Yes.
B: Well now erm, I'm really good at it.
A: Can you do all the musical notes?

B: Yeah.
— text omitted —
B: Guess what Kirsty was doing when
A: What?
B: we was just practising for recorders?
A: What?
B: She was going like this, and the music was on, she put her feet out and she put the music on her feet.
A: Oh well.

Finally, Type 4 conversations are also relatively common (308 texts). This cluster is relatively unmarked on Dimensions 1 and 2, but these conversations are extremely narrative in their Dimension 3 characterization. Text Sample 4 illustrates these characteristics. Note that these conversations are not necessarily extended stories (although some of them are). Rather, as in the present case, these conversations can be constructed out of extended discussion of past events (with frequent past tense verb phrases, 3rd person pronouns, and communication verbs — especially *said* in this conversation), often coupled with commentary on their immediate relevance.

Text Sample 4. Conversation from Cluster 4: Narrative

A: I've just explained that to him. And he said he didn't know that, that he would get hold of Sen and ring me first thing, thing in the morning . — er, to tell me why Sen hasn't paid. He's got the invoice and everything. I said well you've sent us twenty thou= . — I said there is no VAT on it which it should be! Deary me! He says. Has he got the invoice? I said yes. And I said, we've been having, having the invoice outstanding since October at two and half thousand pound! I said, you actually owe me six thousand, one hundred and odd! And I said, you must realize I've a small company, and that's, in one respect that I've had to send those conditions because you're failing to meet the agreed thirty days payment!

B: Yeah.

A: And I said it's not on! I said we couldn't survive like that. And he said, well would you like to carry on with the contract? I said we're too far committed now to, I says to back out. I said, you know, we can't back out at this stage. And I said, but I said if there isn't the payments of the invoice when they are sent . — then . — you know, we've go= you've gotta look at it. So that invoice wants -

B: Doing. Yeah.

A: it wants doing and sending, and put in i= put twenty eight days on.

B: Yeah.

A: Had to be paid, it can't be paid by the twenty eighth it's er . — you know — well if I could've got hold of David or er, Andrew <name>, I was gonna give Andrew <name> a right bollocking for just pushing it in and he should've sent it to er, Michael <name>, Michael <name>'s just got it shoved in front of his nose a= in Edinburgh. He's just gone in to see if everything's alright at Edinburgh . — and of course, that's why he's had to report for that. Which was fair play to him,

but bloody Andrew should have told him! It's agreed, the system of stage payments, it's all written to him.

B: And you've just spoken to him have you?

A: I've just spoken to Michael <name>. Michael's great!

B: And you, you . — so he understands after he's sent you this?

A: What?

B: What's going on.

A: Yeah. Because yo= di= I said I had to send that agreement because you're failing to meet the standard agreement, you're not paying within the twenty eight days or the thirty days!

B: Mm.

A: I said I've got an invoice outstanding for October, and I said I can't afford to do that! He said, I realize that. Then he said, we would want you to do that work he said, because you've got a good reputation. — It makes, you know, if we — we're not gonna go bust just to get twelve months bloody work out of him on a service contract! You know, but i= if we couldn't, if they . — we= . — as they said, if they wanted that money back tomorrow we could only give them half that money back because of what we've got.

6. Conclusion

The three dimensions identified by this factor analysis of a conversational corpus are surprisingly similar to the dimensions of variation found in the earlier MD analysis of general spoken and written registers (Biber, 1988). Both analyses have a dimension that reflects the distinction between involved/interactive versus informational discourse; both analyses have a narrative dimension; and both have dimensions related to the expression of stance. The large-scale MD analyses of spoken and written registers in Somali and Korean similarly identified dimensions associated with these functions; composed of similar kinds of linguistic features.

Even more surprisingly, several MD analyses of restricted discourse domains have identified dimensions with similar formal and functional correlates (compare, for example, Reppen's (1994, 2001) analysis of elementary school registers with White's (1994) analysis of job interview registers). The fact that similar dimensions are found to be basic even in a corpus restricted to conversation suggests that these might be candidates for universal parameters of variation.

Comparing the present analysis to previous MD studies provides two complementary perspectives on the characteristics of conversation. In comparison to the full range of spoken and written registers, conversation is distinctive in being extremely interactive, involved, focused on the immediate context and personal stance, and constrained by real-time production circumstances. However, when conversation is considered on its own terms, we discover systematic patterns of variation among conversational texts (see also Carter and McCarthy, 1997; McCarthy, 1998; Quaglio, 2004; Quaglio and Biber, to appear). Interestingly, the present analysis indicates that the major parameters of variation internal to conversation are a mirror image to the dimensions of variation that distinguish among spoken and written registers.

References

Biber D. (1985). Investigating macroscopic textual variation through multi-feature/multi-dimensional analyses. *Linguistics*, vol. (23): 337-60.

Biber D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, vol. (62): 384-414.

Biber D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber D. (1989). A typology of English texts. *Linguistics*, vol. (27): 3-43.

Biber D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*, vol. (5): 257-269.

Biber D. (1992). On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes*, vol. (15) : 133-163. (Reprinted in Conrad and Biber (Eds) (2001): 215-240.)

Biber D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics*, vol. (19): 219-241.

Biber D. (1994). An analytical framework for register studies. In Biber D. and Finegan E. (Eds), *Sociolinguistic perspectives on register*. Oxford University Press: 31-56.

Biber D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

Biber D. (2001). Dimensions of variation among 18th century speech-based and written registers. In Conrad S. and Biber D. (Eds): 200-214.

Biber D. (2003), Variation among university spoken and written registers: A new multi-dimensional analysis. In Meyer C. and Leistyna P. (Eds), *Corpus analysis: Language structure and language use*. Rodopi.

Biber D. and Finegan E. (2001). Diachronic relations among speech-based and written registers in English. In Conrad S. and Biber D. (Eds): 66-83.

Biber D., Johansson S., Leech G., Conrad S. and Finegan E. (1999). *The Longman grammar of spoken and written English*. Longman.

Brown P. and Fraser C. (1979). Speech as a marker of situation. In Scherer K.R. and Giles H. (Eds), *Social markers in speech*. Cambridge University Press: 33-62.

Carter R. and McCarthy M. (1997). *Exploring Spoken English*. Cambridge University Press.

Connor-Linton J. (1989). *Crosstalk: A multi-feature analysis of Soviet-American spacebridges*. Ph.D. Dissertation: University of Southern California.

Conrad S. and Biber D. (Eds) (2001). *Variation in English: Multi-Dimensional Studies*. Longman.

Ervin-Tripp S. (1972). On sociolinguistic rules: Alternation and co-occurrence. In Gumperz J.J. and Hymes D. (Eds), *Directions in sociolinguistics*. Holt: 213-250

Halliday M.A.K. (1988). On the language of physical science. In Ghadessy M. (Ed.), *Registers of written English: Situational factors and linguistic features*. Pinter: 162-178

Hymes D. (1974). *Foundations in sociolinguistics: An ethnographic approach*. University of Pennsylvania Press.

Jang S.-Ch. (1998). *Dimensions of spoken and written Taiwanese: A corpus-based register study*. Ph.D. Dissertation. University of Hawaii.

Grabe W. (1987). Contrastive rhetoric and text-type research. In Connor U. and Kaplan R.B. (Eds.), *Writing across languages: Analysis of L2 text*. Addison-Wesley: 115-138.

Kanoksilapatham B. (2003). *A Corpus-based Investigation of Biochemistry Research Articles: Linking Move Analysis with Multidimensional Anal*ysis. Ph.D. Dissertation. Georgetown University.

McCarthy M. (1998). *Spoken Language and Applied Linguistics*. Cambridge University Press.

Quaglio P. (in preparation). *Conversation and TV Dialogue: A Corpus-based Study of NBC's Friends*. Ph.D. Dissertation. Northern Arizona University.

Quaglio P. and Biber D. (to appear). The grammar of conversation. In McMahon A. and Aarts B. (Eds), *The Handbook of English Linguistics*. Blackwell.

Reppen R. (1994). *Variation in elementary student writing*. Ph.D. Dissertation. Northern Arizona University.

Reppen R. (2001). Register variation in student and adult speech and writing. In Conrad S. and Biber D. (Eds): 187-199.

White M. (1994*). Language in job interviews: Differences relating to success and socioeconomic variables*. Ph.D. Dissertation. Northern Arizona University.

Appendix A.
List of grammatical, syntactic, lexico-grammatical, and semantic features identified by the Biber Tagger

1. Pronouns and pro-verbs

first person pronouns
second person pronouns
third person pronouns (excluding it)
pronoun it
demonstrative pronouns (this, that, these, those as pronouns)
indefinite pronouns (e.g., anybody, nothing, someone)
pro-verb do

2. Reduced forms and dispreferred structures

contractions
subordinator that deletion (e.g., I think [that/0] he went)
stranded prepositions (e.g., the candidate that I was thinking of)
split auxiliaries (e.g., they were apparently shown to …)

3. Prepositional phrases

4. Coordination

phrasal coordination (NOUN and NOUN; ADJ and ADJ; VERB and VERB; ADV and ADV)
independent clause coordination (clause initial and)

5. WH-Questions

6. Lexical specifity

type/token ratio
word length

7. Nouns

nominalizations (ending in –tion, -ment, -ness, -ity)
nouns

7a. Semantic categories of nouns

animate noun (e.g., teacher, doctor, employee …)
cognitive noun (e.g., fact, knowledge, understanding …)
concrete noun (e.g., rain, sediment, modem …)
technical/concrete noun
quantity noun (e.g., date, energy, minute …)
place noun (e.g., habitat, room, ocean …)
group/institution noun (e.g., committee, bank, congress …)
abstract/process nouns (e.g., application, meeting, balance …)

8. Verbs

8a. Tense and aspect markers

past tense
perfect aspect verbs
non-past tense

8b. Passives

agentless passives
by passives

8c. Modals

possibility modals (can, may, might, could)
necessity modals (ought, must, should)
predictive modals (will, would, shall)

8d. Semantic categories of verbs

be as main verb
activity verb (e.g., smile, bring, open)
communication verb (e.g., suggest, declare, tell)
mental verb (e.g., know, think, believe)
causative verb (e.g., let, assist, permit)
occurrence verb (e.g., increase, grow, become)
existence verb (e.g., possess, reveal, include)
aspectual verb (e.g., keep, begin, continue)

8e. Phrasal verbs

intransitive activity phrasal verb (e.g., come on, sit down)
transitive activity phrasal verb (e.g., carry out, set up)
transitive mental phrasal verb (e.g., find out, give up)
transitive communication phrasal verb (e.g., point out)
intransitive occurrence phrasal verb (e.g., come off, run out)
copular phrasal verb (e.g., turn out)
aspectual phrasal verb (e.g., go on)

9. Adjectives

attributive adjectives
predicative adjectives

9a. Semantic categories of adjectives

size attributive adjectives (e.g., big, high, long)
time attributive adjectives (e.g., new, young, old)
color attributive adjectives (e.g., white, red, dark)
evaluative attributive adjectives (e.g., important, best, simple)
relational attributive adjectives (e.g., general, total, various)
topical attributive adjectives (e.g., political, economic, physical)

10. Adverbs and adverbials

place adverbials
time adverbials

10a. Adverb classes

conjuncts (e.g., consequently, furthermore, however)
downtoners (e.g., barely, nearly, slightly)
hedges (e.g., at about, something like, almost)
amplifiers (e.g., absolutely, extremely, perfectly)
emphatics (e.g., a lot, for sure, really)
discourse particles (e.g., sentence initial well, now, anyway)
other adverbs

10b. Semantic categories of stance adverbs

non-factual/manner-of-speaking adverbs (e.g., frankly, mainly, truthfully)

attitudinal adverbs (e.g., surprisingly, hopefully, wisely)
factual adverbs (e.g., undoubtedly, obviously, certainly)
likelihood adverbs (e.g., evidently, predictably, roughly)

11. Adverbial subordination

causative adverbial subordinator (because)
conditional adverbial subordinator (if, unless)
other adverbial subordinator (e.g., since, while, whereas)

12. Nominal post-modifying clauses

that relatives (e.g., the dog that bit me, the dog that I saw)
WH relatives on object position (e.g., the man who Sally likes)
WH relatives on subject position (e.g., the man who likes popcorn)
WH relatives with fronted preposition (e.g., the manner in which he was told)
past participial postnominal (reduced relative) clauses (e.g., the solution produced by this process)

13. *That* complement clauses

13a. That clauses controlled by a verb (e.g., we predict that the water is here)

non-factual/communication verb (e.g., imply, report, say, suggest)
attitudinal verb (e.g., anticipate, expect, prefer)
factual/mental verb (e.g., demonstrate, know, realize, show)
likelihood/mental verb (e.g., appear, hypothesize, predict, think)

13b. That clauses controlled by an adjective (e.g., it is strange that he went there)

attitudinal adjectives (e.g., good, advisable, paradoxical)
likelihood adjectives (e.g., possible, likely, unlikely)
other adjectives

13c. That clauses controlled by a noun (e.g., the proposal that he put forward was accepted)

non-factive noun (e.g., comment, proposal, remark)
attitudinal noun (e.g., hope, reason, view)
factive noun (e.g., assertion, observation, statement)
likelihood noun (e.g., assumption, implication, opinion)

14. WH-clauses

15. To-clauses

15a. To-clauses controlled by a verb (e.g., He offered to stay)

speech act verb (e.g., urge, report, convince)
cognition verb (e.g., believe, learn, pretend)
desire/intent/decision verb (e.g., aim, hope, like, prefer, want)
modality/cause/effort verb (e.g., allow, leave, order)
probability/simple fact verb (e.g., appear, happen, seem)

15b. To-clauses controlled by an adjective

certainty adjectives (e.g., prone, due, apt)
ability/will adjectives (e.g., competent, hesitant)
personal affect adjectives (e.g., annoyed, nervous)
ease/difficulty adjectives (e.g., easy, impossible)
evaluative adjectives (e.g., convenient, smart)

15c. To-clauses controlled by a noun (e.g., agreement, authority, intention)

Statistical Analysis of Text in Educational Measurement

Claudia Leacock

Educational Testing Service – Rosedale Road – Princeton, NJ 08541 – USA
cleacock@ets.org

Abstract

This paper describes tools developed at Educational Testing Service which use statistical modeling of textual corpora to provide automated assessments of student responses. Details are given for three of these systems: *Critique* Writing Analysis Tools for providing feedback to students, *e*-rater for assigning a holistic score to student essays, and *c*-rater, which scores responses to content-based, short-answer test or chapter review questions.

Keywords: automated scoring, *e*-rater, *c*-rater.

1. Introduction

At Educational Testing Service (ETS), a team of linguists and computer scientists that specializes in natural language processing have developed a variety of tools for instruction and for educational measurement. The goal of our work is to assist students and teachers by providing feedback on the form, quality, organizational structure, and content of writing. It is intended to be an aid, *not a replacement*, for classroom instruction. By providing automated feedback and essay evaluation, our tools ease the instructor's load, thereby enabling the instructor to give students more practice writing essays and answering test questions.

This paper describes three systems developed and deployed by ETS: *Critique* Writing Analysis Tools, *e*-rater, and *c*-rater. The *Critique* Writing Analysis Tools detect numerous errors in grammar, usage, and mechanics. They also highlight undesirable stylistic elements and provide information about essay-based discourse structure. *E*-rater assigns an overall or holistic score to an essay based on the kinds of criteria that human readers are asked to use in evaluating writing on standardized tests, such as the Test of English as a Foreign Language (TOEFL). *C*-rater scores responses to content-based questions by comparing each one to a gold-standard model of the correct answer. Although these applications vary in their methods and goals, they have much in common. Each was built from a training corpus in which features were extracted from examples, many of which had previously been categorized by human judges. The features were then used to build a model through statistical learning techniques (for *Critique* and *e*-rater) or manually (for *c*-rater) in order to categorize new student responses.

2. Grammar, Usage and Mechanics

The *Critique* Writing Analysis Tools identify errors in grammar, usage and mechanics. Grammar errors that are identified include subject-verb agreement (**A popular Mexican food are tacos*), ill-formed verbs (**It is must be miserable*), and the improper use of pronouns (**Them are my two reasons*). Usage errors include determiner noun agreement errors (**those*

problem) and faulty comparatives (**most strangest*). Finally, mechanics errors include punctuation errors such as a missing apostrophe (*the teachers book*) or hyphen (*the better known name*), or repetition of, for example, determiners (**the another town*).

A corpus-based, statistical approach is used to detect these violations of English grammar. The system is trained on a large corpus of edited text, from which it extracts and counts *bigrams* that consist of sequences of adjacent word and part-of-speech pairs. The system then searches student essays for bigrams that occur *much less often* than is expected based on the corpus frequencies.

The expected frequencies come from a model of English that is based on 30-million words of newspaper text. Every word in the corpus is tagged using a part-of-speech tagger (Ratnaparkhi, 1996) that has been trained on student essays. This tag set is subsequently enriched to include information about case and definiteness. For example, the sequence "I dropped the pencil" would be tagged as: I_PPSS dropped_VBD the_ATI pencil_NN. The tag PPSS indicates that "I" is a singular subject pronoun, VBD the past form of a verb, ATI a singular definite determiner and NN a singular noun. Bigrams are created from adjacent sequences of part-of-speech tags and of function words. The bigrams generated from the sequence are: I_VBD, PPSS_VBD, VBD_the, VBD_ATI, the_NN, ATI_NN.

To detect violations of English grammar, the system compares observed and expected frequencies in the corpus. The statistical methods that the system uses are commonly used by researchers to detect combinations of words that occur with greater frequency than would be expected if the words were independent. These methods are typically used to find technical terms or collocations. We use the measures for the opposite purpose – to find combinations that occur *less often* than expected, and therefore might be evidence of a grammatical error (Chodorow and Leacock, 2000).

The system uses two complementary methods to measure association: pointwise mutual information and the log-likelihood ratio. Pointwise mutual information gives the direction of association (whether a bigram occurs more often or less often than expected based on the frequencies of its parts). However, mutual information is known to be unreliable when the data are sparse, while the log-likelihood ratio performs better with sparse data. The log-likelihood ratio gives the likelihood that the elements in a sequence are independent (we are looking for non-independent, dis-associated words), but it does not tell whether the sequence occurs more often or less often than expected. By using both measures, we get both the direction and the strength of association, and performance is better than it would otherwise be when data are limited.

Of course, a model based on adjacent pairs cannot capture English grammar. For this reason, we have developed rule-based filters to allow for low probability sequences that are, in fact, grammatical. For example, in the sequence *these pencil erasers*, where the singular noun *pencil* is part of a plural compound noun, the error message is suppressed.

3. Confusable Words

Some of the most common errors in writing are due to the confusion of homophones, words that sound alike but are spelled differently. Martin Chodorow implemented a system to detect errors among *their/there/they're, its/it's, write/right* and hundreds of other homophone sets. For the most common homophones, we extracted 10,000 training examples of correct usage from newspaper text and built a representation of the local context in which each confusable word occurs. The context we use is a five word window : the two words and part-of-speech

tags that precede the confusable word, and the two that follow it. For example, a context for *right* might be "*find the right person to*", consisting of a verb and determiner that precede the homophone, and a noun that follows it. For *write*, an example of a local context is "*they will write the script*", where *write* is preceded by a subject pronoun and modal verb, and followed by a determiner and noun.

Sometimes one or both words in a confusable word pair are so rare that a large training set cannot be assembled from the corpora available to us. One example is the verb *purl* in the pair *purl/pearl*. In this case, Chodorow has developed generic representations. The generic local context for nouns consists of all the part-of-speech tags found in the two positions preceding each noun and in the two positions that follow it in a large textual corpus. Generic local contexts are created for verbs, adjectives, adverbs, and so on. These serve the same role as the word-specific representations built for more common homophones. Thus, *purl* would be represented as a generic verb and *pearl* as a generic noun.

The frequencies found in training are used to estimate the probabilities that particular words and parts of speech will be found at each position in the local context. When a confusable word is encountered in an essay, *Critique* uses a Bayesian classifier (Golding 1995) to select the more probable member of its homophone set, given the local context in which it occurs. If this is not the word that appears in the essay, then the system highlights it as an error and suggests the more probable homophone. For example, when the system encounters *write* in "We have the *write* to express our opinions", it is highlighted and a pop-up window suggests that it be changed to *right.*

For reporting errors that are detected using bigrams and errors caused by the misuse of confusable words, we have chosen to err on the side of precision over recall. That is, we would rather miss an error than tell the student that a well-formed clause is ill-formed. A minimum threshold of 90% precision was set in order for a bigram error or confusable word set to be included in the writing analysis tools.

Since the threshold for precision is between 90-100%, the recall varies from bigram to bigram and confusable word set to confusable word set. To estimate recall, 5,000 sentences were annotated to identify specific types of grammatical errors. For example, the writing analysis tools correctly detected 40% of the subject-verb agreement errors that the annotators had identified and 70% of the possessive marker errors. Errors in the use of confusable word were detected 71% of the time.

4. Elements of Style

Critique also looks at larger domains to give some stylistic information about the essay as a whole, such as the number of words and number of sentences. It points out stylistic constructions and forms that are usually, but not always, considered to be undesirable, such as the use of passive sentences and of overly-long sentences. Of most interest to us, since it is a real problem in student essays, is the identification of highly repetitious use of words in an essay.

In order to create the training corpus, two writing experts identified which words in an essay are repeated or overused so much that the repetition interferes with a smooth reading of the essay. Since this is a very subjective judgment – what is irritating to one reader may seem fine to another – repetitious word use is not presented as being an error. Instead, *Criterion* highlights repetitiveness and lets the student judge whether the essay would be improved with revision.

The system uses a machine learning approach to finding excessive repetition. It was trained on a corpus of 300 essays in which two judges had labeled the occurrences of overly repetitious words. Seven features were found to reliably predict which word(s) should be labeled as being repetitious. These include the word's total number of occurrences in the essay, its relative frequency in the essay and in a paragraph, its length in characters, and the average distance between its successive occurrences. Using these features, a decision-based machine learning algorithm, C5.0, was used to model repetitious word use, based on the human experts' annotations. See Burstein and Wolska (2003) for a detailed description and system results.

Not surprisingly, the two human judges showed considerable disagreement with each other in this task, but each judge was internally consistent. When the repetitious word detection system was trained on data of a single judge, it could accurately model that individual's performance with 95% precision and 90% recall.

5. Discourse Structure

Scoring rubrics for essays typically define an excellent essay as one that "*develops its arguments with specific, well-elaborated support*" and a poor essay as one that "*lacks organization and is confused and difficult to follow*." In order for an essay to be well-structured, it should contain introductory material, a thesis statement, several main and supporting ideas, and a conclusion.

The organization and development module identifies these discourse components in each essay. In order to do this, a training corpus was created by two writing experts who annotated a large sample of student essays identifying and labeling each discourse unit in the essays. The discourse analysis component uses a decision-based voting algorithm that takes into account the discourse labeling decisions of three independent discourse analysis systems. Two of the three systems use probabilistic methods, and the third uses a decision-based approach to classify a sentence in an essay as a particular discourse element. For a description of the three classifiers and detailed results, see Burstein *et al.* (2003).

When too few discourse elements are identified, *Critique* offers suggestions to the student about how to improve the essay's discourse structure. For example, if only a single main idea is identified, then the *Critique* will advise the student to incorporate two more main ideas into the essay.

To evaluate the system's performance, it was compared against a baseline algorithm. The baseline algorithm assigns a discourse label to each sentence in an essay based solely on its position within the essay. For example, a baseline algorithm would label the first sentence of every paragraph in the essay as a main point. For the baseline algorithm, the overall precision is 71% and the overall is 70% while the precision and the recall for the discourse analysis system are both 85%.

6. *E*-rater: Essay Scoring

In large assessment programs in the United States, such as TOEFL or the Graduate Management Admissions Test, the student is asked to write an essay-length response to a writing prompt within a limited amount of time. A holistic score is assigned to the essay based not so much on the content of the student's response (there is no correct or incorrect answer), but on how well a student frames that response in the context of an essay. These

holistic scores usually range from 1 for a poorly written essay to 6 for an excellent, well-crafted essay. The holistic scoring rubric bases the essay score on such features as the essay's clarity and persuasiveness, the organization and development of its ideas, the use of sentence variety, the choice of language, and the essay's overall grammatical correctness.

E-rater is the automated essay scoring engine that has been developed and deployed at ETS. *E*-rater version 2.0, which is described here, was invented by Jill Burstein, Yigal Attali and Slava Andreyev and will be deployed in Spring, 2004. The *e*-rater 2.0 feature set is closely aligned to the scoring rubric. For example, two features that are derived from the Writing Analysis Tools' organization and development component directly address the essay's organization and the development of its ideas. Three features are derived from the grammar, usage and mechanics components to assess the grammatical correctness and use of punctuation in the essay. Features that are based on content vector analysis, type/token ratios, and an index of word difficulty evaluate the choice of language. For a full description of *e*-rater 2.0, see Burstein, et al (forthcoming).

To model human scores, *e*-rater needs to combine this homogeneous set of features that correspond to the elements in the scoring rubric. It builds an individual model for each essay question by training on a sample of about 250 essays that two human readers have scored and that represent the range of scores from 1 to 6. *E*-rater uses a multiple regression approach to generate weights for the feature set. The result of training is a regression equation that can be applied to the features of a novel essay to produce a predicted score value. The last step in assigning an *e-rater* score is to convert the continuous regression value to a whole number along the six-point scoring scale.

The performance of *e*-rater 2.0 is evaluated by comparing its scores to those of human judges. This is carried out in the same manner that the scores of two judges are measured during reader scoring sessions for standardized tests such as TOEFL. If two judges' scores match exactly, or if they are within one point of each other on the 6-point scale, they are considered to be in *agreement*. Typical agreement between *e-rater* 2.0 and the human score is approximately 97%, which is comparable to agreement between two human readers.

7. *C*-rater: Scoring for Content

C- rater is an automated scoring engine under development at ETS that is designed to measure a student's understanding of specific content material. Unlike the Writing Analysis Tools and *e*-rater, it measures the student's understanding with little regard for writing skills.

Below is an example of the type of question that c-rater is designed to score. This example is an approximation of a 4th grade math question used in the National Assessment for Educational Progress (NAEP):

> A radio station wanted to determine the most popular type of music among those in the listening range of the station. Would sampling opinions at a country music concert held in the listening area of the station be a good way to do this?
>
> ▢ YES ▢ NO
>
> Explain your answer.

To answer the first part of the question, the student indicates "yes" or "no" by filling in a bubble. *C*-rater can be used to score the second part of the question – the student's explanation.

To create a *c*-rater scoring model, a content expert, such as a test developer or a teacher, needs to develop a scoring guide that breaks down how many score points are assigned to each of the component concepts that make up a correct answer. In the example above, only one concept needs to be identified in order for the student to receive credit – the concept that the sample would be biased. The content expert, working with a set of scored student responses, must identify the variety of ways that the concept can be expressed lexically and syntactically. We call these *gold standard* responses. In this case, most of the students express the concept of bias by saying something like: "They would say that they like country music".

The *c*-rater scoring engine tries to recognize when a response is equivalent to a correct answer. *C*-rater is, in essence, a paraphrase recognizer. In order to recognize paraphrases, it breaks down each student response into its underlying predicate argument structure, resolves the referent of any pronouns in the response, regularizes over morphological variation, matches on synonyms or similar words, and tries to resolve the spelling of unrecognized words. The resulting canonical representations are mapped onto canonical representations of the gold standard responses.

In the current version of *c*-rater, the module that maps between the canonical representation of a gold standard answers and that of a student response is rule driven – *not* statistical. The rule-driven approach allows for good accuracy. In a pilot study with the state of Indiana, *c*-rater was used to score seven 11th grade reading comprehension items. On average, *c*-rater had exact agreement with human scorers 84% of the time. (See Leacock and Chodorow, 2003b for a full description of *c*-rater and the pilot study.) However, the scoring engine is unable to give an indication of its confidence about the score. If *c*-rater could generate a reliable measure of confidence, then it could be used to grade those responses that it is confident about and an instructor would only need to grade the responses on which system confidence is low.

Thomas Morton is currently developing a statistical version of *c*-rater that gives a score as well as a probability indicating the system's confidence that the score is correct. This version trains a maximum entropy model on a corpus of student responses, where the mapping to one or more corresponding gold standard sentences whose meaning they capture was manually annotated. The features used to do this are based on lexical similarity as well as structural similarity. The lexically-based features include exact string match, head word match, and synonymous words. The structural features are based on the labels of the constituents where matching arguments are found. For example, the model can learn that a subject can typically be matched to an object in a passive sentence. Thus the model learns which syntactic variations maintain meaning and which ones do not. Since this model is based on syntactic variations and lexical similarity, and not the actual lexical items found in the corpus, it can be used for any question that exhibits the same types of syntactic variation that are found in the training data.

Both versions of *c*-rater generate a score based on the similarity of the response to the gold standard answer. This is done by comparing each gold standard sentence to each sentence in the response, and a score is assigned based on which concepts are present in the response. To generate a measure of confidence in the score, the machine learning version keeps a ranked set of the most likely mappings between a response and the gold sentences. Each of these sets of mappings is scored and the weight of the set is used to produce a distribution of scores.

8. Conclusion and Future Directions

In developing tools for instruction and for educational measurement we have systematically exploited the information that can be found in annotated and unannotated textual corpora.

With *e*-rater, the annotation consists of a holistic score. In the Writing Analysis Tools, the annotations either identify the elements of discourse in an essay or identify overly-repetitious word use. Statistical classifiers train on the annotated corpora to learn how to annotate new test responses. In *c*-rater, a corpus of aligned sentences is used to generate a model of meaning-preserving variations. We have also exploited large unannotated corpora of published, well-formed text to build a model of well-formed English against which individual student responses are compared in order to identify errors in grammar, usage, and mechanics.

We are currently expanding the functionality of the Writing Analysis Tools, *e*-rater and *c*-rater. For example, we are implementing the detection of grammatical errors that are important to specific native language groups (Han *et al.*), such as identifying when a determiner is missing (a common error among native speakers of East Asian languages and of Russian) or when the wrong preposition is used. The current organization and development module of the Writing Analysis Tools identifies discourse elements but does not evaluate their quality. We are extending the analysis of discourse so that the expressive quality of each discourse element can also be assessed. As previously mentioned, we also plan to replace the c-rater module that assigns a score with a statistically based module that assigns a score along with a confidence measure.

Acknowledgements: I am grateful to Martin Chodorow, Ray C. Dougherty and Thomas Morton for helpful comments and suggestions. Any opinions expressed here are those of the authors and not necessarily of the Educational Testing Service.

References

Burstein J., Chodorow M. and Leacock C. (forthcoming). Automated essay evaluation: The *Criterion*^{SM} Online Service. *AI Magazine.*

Burstein J., Marcu D. and Knight K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems: Special Issue on Natural Language Processing*, vol. (18/1): 32-39.

Burstein J. and Wolska M. (2003). Toward evaluation of writing style: Overly repetitious word use in student writing. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*. Budapest, Hungary.

Chodorow M. and Leacock C. (2000). An unsupervised method for detecting grammatical errors. *Proceedings of the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics*: 140-147.

Golding A. (1995). A Bayesian hybrid for context-sensitive spelling correction. In *Proceedings of the Third Workshop on Very Large Corpora*, Cambridge, MA: 39-53.

Han N-R., Chodorow M. and Leacock C. (2004). Detecting errors in English article usage with a maximum entropy classifier trained on a large, diverse corpus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. Lisbon, Portugal.

Leacock C. and Chodorow M. (2003a). Automated grammatical error detection. In Shermis M.D. and Burstein J. (Eds), *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Lawrence Erlbaum.

Leacock C. and Chodorow M. (2003b). C-rater: Automated scoring of short answer questions. *Computers and the Humanities*, vol. (37/4).

Ratnaparkhi A. 1996. A maximum entropy part-of-speech tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*. University of Pennsylvania.

Étude de la stabilité des valeurs propres de l'AFC d'un tableau lexical au moyen de procédures de rééchantillonage

Ramón Álvarez[1], Mónica Bécue[2], Olga Valencia[3]

[1]Universidad de León. León – Spain – dderae@unileon.es
[2]Universidad Politécnica de Cataluña. Barcelona – Spain – monica.becue@upc.es
[3]Universidad de Burgos. Burgos – Spain – oval@ubu.es

Abstract

This paper studies the external stability of eigenvalues issued from Correspondence Analysis of a lexical table, by means of resampling. The problems of lack of stability are more likely to happen in this particular kind of contingency tables, the lexical tables, due to the sparsity of the matrix and other characteristics.
The non-parametric Bootstrap applied in combination with factorial methods which use a Singular Value Decomposition (SVD) algorithm, as Correspondence Analysis does, results in "extra variability" of the statistics involved, including the eigenvalues. Reflections, permutations and rotations of axes may take place in the Bootstrapped samples, that make comparisons between each simulated configuration and the original one quite difficult to understand and senseless in some way. But the whole variability observed doesn't measure the real external stability degree. To study in depth the latter, subsequent corrections must be applied to the results issued from the Total Bootstrap. This may be achieved by a suitable Orthogonal Procustes Rotation.

Résumé

Cet article étudie la stabilité externe des valeurs propres fournies par l'Analyse de Correspondances d'un tableau lexical au moyen de rééchantillonnage. Les problèmes de stabilité des résultats sont plus fréquents dans ce type de tableau de contingence quasi-vide ou creux.
Le Bootstrap non paramétrique est appliqué en combinaison avec les méthodes factorielles qui emploient un algorithme de décomposition en valeurs singulières (SVD), comme l'Analyse de Correspondances, ce qui entraîne une variabilité supplémentaire des statistiques impliquées, y compris les valeurs propres. On peut avoir à faire face à des réflexions, permutations et rotations des axes dans les analyses effectuées sur les échantillons bootstrap, ce qui rend difficile les comparaisons entre chaque configuration simulée et l'originale. Mais la variabilité totale observée ne mesure pas le vrai degré de stabilité externe. Pour vraiment étudier la stabilité, il faut appliquer quelques corrections aux résultats, en particulier en applicant une rotation orthogonale de type Procustes.

Keywords: eigenvalues, correspondence analysis, bootstrap, lexical tables, Procustes analysis

1. Introduction

Le travail présenté ici[1] considère que le caractère exploratoire d'une technique n'implique pas que ses résultats ne soient pas soumis à un certain type de « contrôle de qualité ». En ce sens, la stabilité des résultats fournis par les analyses factorielles exploratoires a été étudiée par divers auteurs. Lebart (1998) étudie la portée et la validité des résultats des Analyses en Composantes Principales (ACP), de l'Analyse Factorielle des Correspondances (AFC) et de

[1] Travail développé grâce au Projet « Analyse Statistique des Enquêtes aux Jeunes Juges » (SEC2001-2581-C02-02, Ministère espagnol pour la Science et la Technologie).

l'Analyse en Correspondances Multiples (ACM) en introduisant certaines perturbations dans la matrice initiale et en suggérant des outils probatoires de diverses natures.

Notre intérêt est centré sur la stabilité externe, ce qui implique la nécessité de vérifier la stabilité face aux fluctuations de l'échantillonnage. En suivant les suggestions de Lebart et d'autres auteurs mentionnés dans la suite de cet article, nous considérons le rééchantillonnage comme une option adéquate pour analyser le comportement de statistiques complexes, comme sont les valeurs propres fournies par une AFC. L'AFC appliquée à des tableaux de contingence lexicaux est un cas particulier qui présente des caractéristiques propres, puisque ces tableaux sont plus propices à l'instabilité.

La procédure de rééchantillonnage choisie est le Bootstrap non paramétrique (Efron, 1993), à cause de son caractère de rééchantillonnage basé sur les données sans hypothèses formelles de travail. Son application pratique requiert cependant une attention spéciale quand il est combiné avec des techniques factorielles, comme l'AFC, qui font usage d'un algorithme de Décomposition en Valeurs Singulières. Cet algorithme, comme l'indiquent Milan et Whittaker (1995), produit une variabilité « supplémentaire » non imputable aux fluctuations du rééchantillonnage, variabilité qui peut affecter la perception initiale du degré de stabilité externe des résultats mais ne doit pas être pris en considération dans son évaluation finale.

La section 2 précise l'objectif de l'étude ; la section 3 présente la méthodologie conçue pour aborder le problème de la stabilité des valeurs propres dans l'AFC d'un tableau lexical ; la section 4 décrit l'exemple choisi (tableau lexical) tandis que la section 5 présente les résultats obtenus ; finalement, la section 6 offre quelques conclusions.

2. Objectif du travail

Le but de ce travail est l'étude de la stabilité externe des valeurs propres de l'AFC d'un tableau lexical au sein d'une étude globale qui inclut aussi l'analyse des coordonnées, des contributions absolues et relatives et des vecteurs propres.

Les tableaux lexicaux sont un type spécial de tableaux de contingence plus enclins à présenter des problèmes d'instabilité dûs à leur nature. En effet, il s'agit de matrices creuses, avec une grande différence entre le nombre de lignes (unités lexicales) et le nombre de colonnes (réponses regroupées par catégories, réponse individuelle ou document textuel classique), dans lesquelles abondent les fréquences marginales très faibles (ce qui concerne tout particulièrement les unités lexicales). D'autre part, l'inertie totale est généralement distribuée presque uniformément entre les axes produits par l'AFC, ce qui entraîne une grande proximité entre les inerties principales ou les valeurs propres. Tout cela est généralement source de problèmes, tant dans la génération des simulations bootstrap elles-mêmes comme dans l'interprétation des facteurs et la représentation graphique des plans factoriels obtenus. Ces difficultés peuvent conditionner la validité des résultats. Dans ce travail, nous centrons notre attention sur la stabilité des valeurs propres.

Nous utilisons une procédure de rééchantillonnage en trois phases :

a. La génération d'un nombre B élevé d'échantillons simulés (B=5000), par Bootstrap non paramétrique.

b. L'application de l'AFC à chacun des échantillons Bootstrap construits (Bootstrap Total).

c. La comparaison des B ensembles de statistiques, associées à chacune des AFC effectuées, avec les statistiques obtenues par l'AFC appliquée à la matrice originale, en particulier comparaison des valeurs propres.

Une comparaison directe des statistiques fournies par le Bootstrap Total d'une AFC avec les statistiques originales peut conduire à une évaluation erronée du degré de stabilité externe, puisque les configurations des échantillons Bootstrap peuvent différer pour des causes « apparentes » et/ou pour des causes « réelles » :

– Nous appelons « apparentes » les différences provoquées par la combinaison du rééchantillonnage Bootstrap avec les méthodes qui, comme les méthodes factorielles, utilisent un algorithme de décomposition en valeurs singulières (SVD) (réflexions de certains facteurs, permutations arbitraires entre des facteurs de rang voisin, changements d'orientation des sous-espaces ou simplement changements d'échelle). Ces modifications des sous-espaces ne supposent pas une réelle modification des distances relatives entre les points et, pour cela, on considère qu'il n'y a pas de véritables différences entre les configurations. Ce pourquoi cette variabilité « supplémentaire » ne doit pas être considérée comme un signe d'instabilité.

– On considère « réelles » les différences causées par les fluctuations de rééchantillonnage. Ces différences doivent servir à évaluer la stabilité externe des résultats. On utilise l'Analyse Procustes Orthogonal pour effectuer une rotation qui permette de superposer chaque configuration bootstrap à la configuration correspondant à l'échantillon original de manière optimale, c'est-à-dire, en obtenant la réduction maximale de la variabilité apparente, et en permettant une comparaison plus adéquate des configurations, a partir de laquelle il soit possible d'effectuer l'analyse de stabilité externe souhaitée.

3. Méthodologie

La méthodologie que nous proposons effectue la comparaison des configurations et, par conséquent, des valeurs propres de trois manières différentes :

1. Par comparaison directe des résultats fournis par le Bootstrap Total. Cette analyse présente des limitations dues à l'influence de la variabilité « supplémentaire » des statistiques, en particulier des valeurs propres, qui vient du fait que les statistiques ne correspondent pas nécessairement aux axes auxquelles elles sont assignées.

2. Par comparaison des résultats fournis par le Bootstrap Total, mais après une rotation Procustes orthogonale destinée à mettre en rapport les axes, et en reconstruisant les valeurs propres (appelées dans ce qui suit pseudo-valeurs propres) à partir des coordonnées

3. Par l'obtention des pseudo-valeurs propres à partir d'un Bootstrap Partiel

Dans ce qui suit, on détaille la méthodologie correspondant au point 2, qui permet de résoudre certains des inconvénients mentionnés dans la section 2, et on précise le rôle et le fonctionnement de l'Analyse Procustes dans ce contexte.

Le principal problème est de comparer chacune des B configurations obtenues C_1, C_2, ..., C_B, avec la configuration associée à l'échantillon original, C_o. L'Analyse Procustes Orthogonale permet d'effectuer un ajustement de chaque configuration bootstrap C_b à la configuration originale C_o, en appliquant une série d'opérations (translation rigide, rotation rigide et dilatation uniforme) de sorte que soit minimale la somme des carrés des distances entre chaque paire de points correspondant à chacune des deux configurations (Krzanowski, 2001).

– La translation est utilisée pour modifier l'origine de l'espace, habituellement pour le faire coïncider avec l'origine de coordonnées et obtenir ainsi une origine commune pour les configurations comparées. Pour cela, il est usuel de travailler avec des matrices centrées par colonne, mais dans notre cas, la translation est faite en multipliant les coordonnées, tant originales C_o comme Bootstrap C_b, par les fréquences marginales respectives F_o et F_b : (C_{op}= F_o C_o

et $C_{bp}= F_b\ C_b$). F_b est la matrice diagonale des fréquences marginales de la simulation b. Ainsi, la somme des coordonnées pondérées C_{bp} est nulle.

– La rotation implique le calcul d'une matrice de rotation R_b pour chacune des matrices bootstrap pondérées C_{bp}, à fin d'obtenir les distances minimales avec la matrice originale pondérée C_{op}.

– La dilatation suppose l'obtention d'une constante scalaire c_b pour adapter la configuration C_{bp} à la taille de la configuration originale pondérée C_{op}.

De cette manière, on obtient de nouvelles matrices avec les configurations pondérées adaptées $C_{bpR} = c_b\ C_{bp}\ Rot_b$, à partir desquelles, en divisant par les fréquences correspondantes F_b, on définit finalement les matrices de coordonnées après rotation C_{bR}. Ces coordonnées représentent des configurations comparables puisque les opérations précédentes ont réduit le plus possible les effets des éventuels réflexions, permutations et changements d'orientation des axes de chaque sous-espace. Les différences entre les configurations demeurant après ces opérations peuvent être considérées comme des différences réelles.

Une fois déterminées les coordonnées après rotation, il est possible de reconstruire les valeurs propres correspondant à chaque axe, à partir des coordonnées sur ces axes ; on obtient ainsi les « pseudo-valeurs propres », ou valeurs propres reconstruites, mentionnées plus haut. Elles correspondent aux inerties principales des nuages de points adaptés de manière beaucoup plus réelle qu'avant la rotation.

Le calcul des pseudo-valeurs propres peut être effectué ou à partir des coordonnées des lignes après rotation ou bien à partir des coordonnées des colonnes après rotation :

$$\lambda^*_{\alpha b} = \sum_{j=1}^{p} f_{.jb} \varphi^{*\ 2}_{j\alpha b} \quad , \text{ ou bien } \lambda^*_{\alpha b} = \sum_{i=1}^{n} f_{i.b} \psi^{*\ 2}_{i\alpha b}$$

La pseudo-inertie totale d'un échantillon bootstrap est égale à la somme des pseudo-valeurs propres.

4. Tableau lexical

Le questionnaire remis aux juges formés à l'École Supérieure de la Magistrature de Barcelone comprenait une question fermée : "*Comment évaluez-vous la formation obtenue à la Faculté de Droit ?* ", avec cinq modalités de réponse : *très négativement, négativement, moyennement, positivement et très positivement.* La question ouverte : "*Pourquoi ?* " était ensuite posée.
On dispose seulement de 268 réponses, ce qui est peu pour procéder à l'analyse de questions ouvertes et pourrait, en principe, affecter négativement la stabilité des résultats. Le tableau lexical comprend 2086 occurrences, réparties en 114 unités lexicales-ligne (formes-lemmes prononcées avec une fréquence supérieure à 5) et 5 colonnes correspondant aux 5 modalités de la question fermée.

5. Résultats

5.1. *Bootstrap Total du tableau lexical sans rotation*

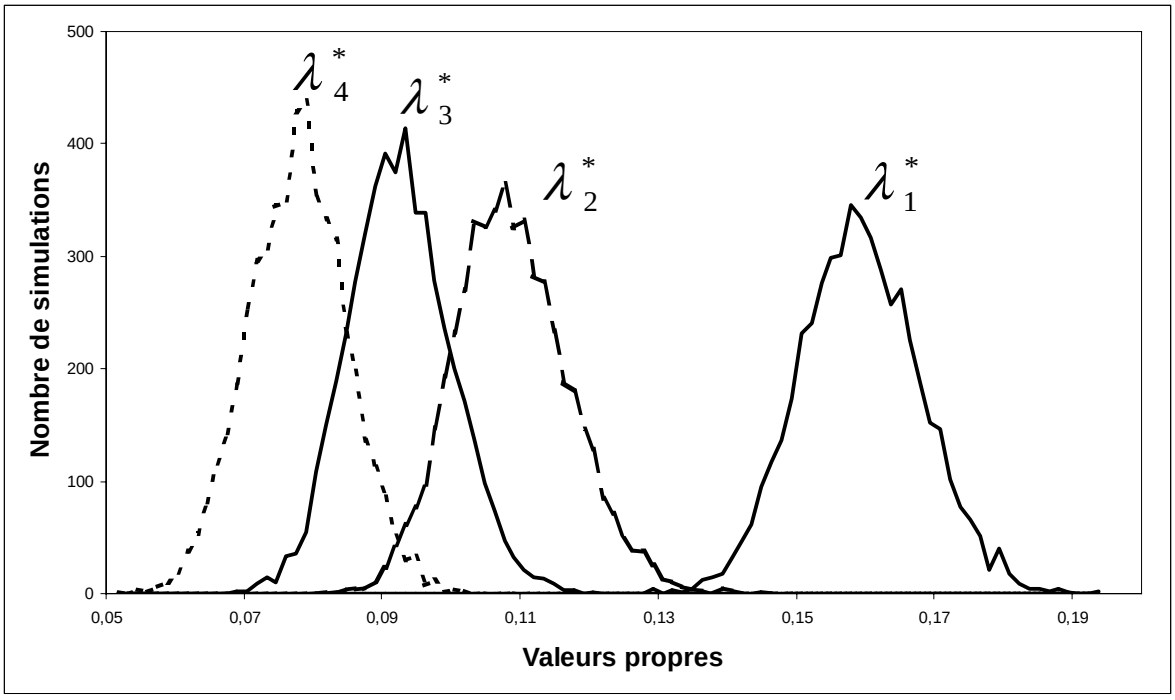
La distribution de l'inertie totale des échantillons bootstrap présente une valeur moyenne (0.4354) très supérieure à l'inertie totale originale (0.3377). Les intervalles percentiles calculés n'incluent pas l'inertie totale d'échantillonnage, ni même en considérant 99 % des simulations. Ceci confirme que le Bootstrap produit une notable augmentation de l'inertie totale et

provoque une perturbation excessive des données initiales. On constate aussi une absence de normalité des inerties totales bootstrap, ce qui indique que la construction d'intervalles de confiance ne peut utiliser le présupposé de la normalité et impose de travailler avec des intervalles percentiles.

Le comportement des inerties principales est spécifié dans le tableau suivant :

	VPROPRE 1	VPROPRE 2	VPROPRE 3	VPROPRE 4
Valeurs propres originaux	**,1378**	**,0797**	**,0643**	**,0559**
Moyenne v.p. Bootstrap	,1584	,1081	,0918	,0770
Écart-type	,0088	,0084	,0074	,0072
Minimum	,1285	,0804	,0678	,0503
Maximum	,1937	,1450	,1196	,1087
Percentiles 0,5	,1364	,0889	,0731	,0590
2,5	,1416	,0927	,0781	,0631
5	,1442	,0951	,0801	,0652
95	,1731	,1226	,1044	,0888
97,5	,1761	,1259	,1070	,0909
99,5	,1810	,1325	,1128	,0960
Test normalité K-S (Sig)	,451	,001	,053	,443
Test norm. Lilliefors (Sig)	,089	,000	,000	,085

Tableau 1. Statistiques des valeurs propres sans rotation



Graphique 1. Distribution des valeurs propres sans rotation

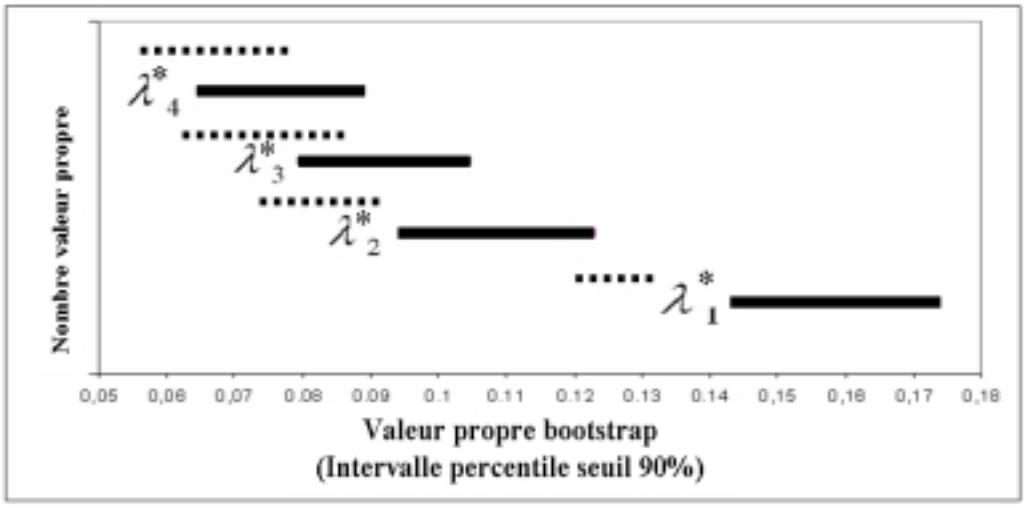
L'interprétation des résultats du Tableau 1 et du Graphique 1 permet d'effectuer quelques observations préliminaires :

– La moyenne des valeurs propres simulées est, dans tous les axes, supérieure à la valeur propre de l'échantillon original : $\overline{\lambda}^*_\alpha > \lambda_\alpha \ \forall\alpha$

– Les inerties principales d'échantillonnage ne sont pas comprises dans les intervalles percentiles au niveau 95 %, ni même au niveau 99 %, sauf l'inertie principale du premier axe. Ceci suggère que la perturbation produite par le bootstrap non seulement donne lieu à une augmentation de l'inertie au niveau global mais à un accroissement de celle-ci dans toutes les dimensions de l'espace factoriel, surestimant systématiquement les véritables inerties principales.

– Au niveau de 95 %, on rejette l'hypothèse de normalité de la distribution des deuxième et troisième valeurs propres, alors que la normalité des première et quatrième valeurs propres ne peut pas être écartée (cette normalité serait cependant rejetée au niveau 90 %).

– Le Graphique 2 représente les intervalles de confiance percentiles au niveau 90 % comme lignes continues. Seul l'intervalle correspondant à la première valeur propre ne recouvre aucun des intervalles correspondant aux autres valeurs propres.



Graphique 2. Intervalles de confiance des percentiles pour les valeurs propres au seuil 90 %

L'examen des résultats directs du Bootstrap Total suggère la stabilité du premier facteur et la possible existence d'instabilité pour le reste des facteurs étant donnée la proximité entre leurs inerties principales. Le Bootstrap Total donne lieu à des estimations non paramétriques des valeurs propres qui se caractérisent tant par leur non-normalité comme par leur biais positif conduisant à surestimer les valeurs-propres.

Ceci nous conduit :

– à rejeter l'utilisation d'intervalles de confiance basés sur la normalité, comme l'intervalle standard avec estimation bootstrap de l'écart type (σ*), et à recourir à des estimations non paramétriques à partir des intervalles percentiles.

– à proposer une correction du biais mentionné, en faisant usage de la même technique Bootstrap pour l'estimer :

$$biais_{\alpha}^{*} = \overline{\lambda_{\alpha}^{*}} - \lambda_{\alpha\,(échantillon)}$$

Le Tableau 2 montre les nouveaux intervalles pour les valeurs propres corrigées, qui comprennent maintenant les valeurs propres originales, construits à partir des intervalles précédents mais en éliminant le biais.

	$\lambda_{éch}$	$\bar{\lambda}^*$	Biais*	$P^*_{5(corr)}$	$P^*_{95(corr)}$	$P^*_{2,5(corr)}$	$P^*_{97,5(corr)}$	$P^*_{0,5(corr)}$	$P^*_{99,5(corr)}$
1	0,1378	0,1584	0,0206	0,1236	0,1525	0,1210	0,1555	0,1158	0,1604
2	0,0797	0,1081	0,0284	0,0667	0,0942	0,0643	0,0975	0,0605	0,1041
3	0,0643	0,0918	0,0275	0,0526	0,0769	0,0506	0,0795	0,0456	0,0853
4	0,0559	0,0770	0,0211	0,0441	0,0677	0,0420	0,0698	0,0379	0,0749

Tableau 2. Intervalles des valeurs propres avec biais corrigés. Percentile au 90 %, 95 % et 99 %.

5.2. *Bootstrap Total du tableau lexical avec rotation*

On considère le tableau de coordonnées C^Y de dimensions ($n+p$, q), juxtaposant la matrice de coordonnées-ligne (n, q) et la matrice de coordonnées-colonne (p, q). On pondère le tableau simulé bootstrap ainsi obtenu ($C^Y_{bp}= F^Y_b\ C^Y_b$) et on le compare au tableau juxtaposant les coordonnées originales (lignes et colonnes) en pondérant chacun des éléments par les fréquences marginales ($C^Y_{op}= F^Y_o\ C^Y_o$).

La rotation Procustes effectuée diminue l'inertie des nuages simulés et offre une pseudo-valeur propre proche de la valeur propre du nuage original. Cela signifie que le caractère excessivement perturbateur du Bootstrap dans l'étude de l'AFC d'un tableau lexical, n'est pas aussi grand que l'on pouvait initialement le penser. Il répond à la nature intrinsèque de la technique Bootstrap, fondamentalement à la variabilité « supplémentaire » produite par sa combinaison avec l'algorithme SVD utilisé par les méthodes factorielles.

En éliminant cette variabilité, l'intervalle percentile au niveau 90 % inclut la valeur originale de l'inertie totale, très proche de la moyenne de celle correspondant aux échantillons bootstrap.

La distribution des pseudo-valeurs propres calculées à partir des coordonnées après rotation, (Tableau 3) permet d'observer que les moyennes des pseudo-valeurs propres sont beaucoup plus proches des valeurs propres originales. Les intervalles percentiles au niveau 90 % incluent les valeurs propres originales λ_2 et λ_3 ; le niveau doit être augmenté jusqu'à 95 % pour λ_4 et jusqu'à 99 % pour λ_1.

La distribution des seconde troisième et quatrième pseudo-valeurs propres ne vérifie pas la normalité, avec un risque de 5 %

Les superpositions entre les intervalles percentiles au niveau 90 % des pseudo-valeurs propres se donnent pour les deuxième, troisième et quatrième axes, tandis que la première valeur propre est encore parfaitement éloignée et différenciée du reste. On ne peut être interpréter, en aucune façon, ces résultats comme un symptôme d'instabilité puisque les trois facteurs mentionnés se sont montrés hautement stables comme l'indiquent les petits intervalles percentiles des valeurs propres correspondantes. Le graphique 2 montre (lignes discontinues) les intervalles de confiance pour les pseudo-valeurs propres.

		VPROPRE 1	VPROPRE 2	VPROPRE 3	VPROPRE 4
Valeurs propres originaux		**,1378**	**,0797**	**,0643**	**,0559**
Moyenne v.p. Bootstrap		,1263	,0830	,0740	,0675
Écart-type		,0048	,0063	,0080	,0070
Minimum		,1086	,0617	,0460	,0437
Maximum		,1456	,1059	,1102	,0935
	Percentiles 0,5	,1143	,0682	,0557	,0506
	2,5	,1170	,0710	,0594	,0543
	5	,1184	,0728	,0615	,0563
	95	,1343	,0935	,0877	,0795
	97,5	,1360	,0954	,0904	,0818
	99,5	,1392	,1003	,0961	,0872
Test normalité K-S (Sig)		,458	,309	,016	,156
Test norm. Lilliefors (Sig)		,092	,033	,000	,005

Tableau 3. Statistiques des pseudo-valeurs propres. Bootstrap Total. Rotation

5.3. *Bootstrap Partiel du tableau lexical*

Le nombre différent de lignes (114) et de colonnes (5) fait que les pseudo-valeurs propres sont différentes si elles sont reconstruites ou bien à partir des lignes ou bien à partir des colonnes.

L'inertie totale des échantillons bootstrap obtenue à partir des pseudo-valeurs propres des colonnes (0.3422) est légèrement supérieure à l'originale (0.3377). Quand on obtient les pseudo-valeurs à partir des lignes la différence est beaucoup plus grande (0.4362). Il y a aussi absence de normalité pour les trois derniers facteurs. Ceci confirme que le Bootstrap produit une augmentation notable de l'inertie totale (surtout pour les lignes dans un tableau lexical), en provoquant une perturbation excessive des données initiales ; on doit donc effectuer la correction de biais proposée au paragraphe 5.1 ou bien une rotation de la configuration bootstrap obtenue avec le Bootstrap Partiel.

Dans le Tableau 4 on observe que les moyennes des pseudo-valeurs propres reconstruites à partir des coordonnées des colonnes sont proches des valeurs propres originales et que l'on rejette l'hypothèse de normalité pour la troisième et la quatrième.

		VPROPRE 1	VPROPRE 2	VPROPRE 3	VPROPRE 4
Valeurs propres originaux		**,1378**	**,0797**	**,0643**	**,0559**
Moyenne v.p. Bootstrap		,1390	,0808	,0656	,0568
Écart-type		,0099	,0085	,0101	,0077
Minimum		,1053	,0529	,0309	,0308
Maximum		,1748	,1123	,1121	,0916
	Percentiles 0,5	,1139	,0602	,0410	,0382
	2,5	,1198	,0645	,0467	,0425
	5	,1228	,0671	,0495	,0445
	95	,1555	,0952	,0826	,0702
	97,5	,1587	,0979	,0858	,0727
	99,5	,1661	,1031	,0931	,0778
Test normalité K-S (Sig)		,492	,683	,111	,033
Test norm. Lilliefors (Sig)		,108	,200*	,002	,000

* Borne inférieure de la signification vraie

Tableau 4. Statistiques des pseudo-valeurs propres colonnes. Bootstrap Partiel. Sans rotation

Mais dans le Tableau 5, les moyennes des pseudo-valeurs propres obtenues à partir des lignes sont assez différentes des valeurs propres originales. Par exemple, pour le facteur 2 la valeur 0.797 est hors des intervalles de confiance. L'absence de normalité dans ce cas affecte les trois derniers facteurs.

	VPROPRE 1	VPROPRE 2	VPROPRE 3	VPROPRE 4
Valeurs propres originaux	**,1378**	**,0797**	**,0643**	**,0559**
Moyenne v.p. Bootstrap	,1566	,1032	,0921	,0843
Écart-type	,0087	,0098	,0119	,0097
Minimum	,1257	,0722	,0514	,0536
Maximum	,1909	,1476	,1470	,1282
Percentiles 0,5	,1355	,0809	,0648	,0623
2,5	,1399	,0850	,0707	,0667
5	,1425	,0878	,0735	,0690
95	,1710	,1199	,1123	,1012
97,5	,1741	,1230	,1166	,1048
99,5	,1790	,1304	,1257	,1123
Test normalité K-S (Sig)	,604	,007	,015	,000
Test norm. Lilliefors (Sig)	,200*	,000	,000	,000

* Borne inférieure de la signification vraie

Tableau 5. Statistiques des pseudo-valeurs propres lignes. Bootstrap Partiel. Sans rotation

Le Tableau 6 montre maintenant les nouveaux intervalles pour les valeurs propres lignes corrigées, intervalles qui comprennent les valeurs propres d'échantillonnage.

	$\lambda_{éch}$	$\bar{\lambda}^*$	Biais*	$P^*_{5(corr)}$	$P^*_{95(corr)}$	$P^*_{2,5(corr)}$	$P^*_{97,5(corr)}$	$P^*_{0,5(corr)}$	$P^*_{99,5(corr)}$
1	0,1378	,1566	0,0188	0,1237	0,1522	0,1211	0,1553	0,1167	0,1602
2	0,0797	,1032	0,0235	0,0643	0,0964	0,0615	0,0995	0,0574	0,1069
3	0,0643	,0921	0,0278	0,0457	0,0845	0,0429	0,0888	0,0370	0,0979
4	0,0559	,0843	0,0284	0,0406	0,0728	0,0383	0,0764	0,0339	0,0839

Tableau 6. Intervalles des pseudo-valeurs propres lignes avec correction du biais. Percentile au 90 %, 95 % et 99 %. Bootstrap Partiel. Sans rotation

6. Conclusions

L'application du Bootstrap non paramétrique est fréquente dans les études de stabilité des résultats des méthodes factorielles exploratoires, en particulier pour l'analyse des coordonnées, ce qui peut être effectué avec un Bootstrap Partiel.

L'inertie totale des tableaux lexicaux obtenus par rééchantillonnage est plus grande que l'originale. Même si on utilise un Bootstrap Partiel, les coordonnées (surtout des lignes-formes) sont surestimées, comme on a pu le vérifier avec le calcul des pseudo-valeurs à partir des coordonnées lignes ou des colonnes.

Le fait que l'inertie des analyses répliquées augmente suggère que les coordonnées bootstrap « vraies » sont plus petites que celles obtenues avec la simulation. Cela conduit à effectuer une correction des pseudo-valeurs propres et à corriger postérieurement les coordonnées, ou à chercher un autre type de simulation moins perturbateur qu'un bootstrap pur, ou à effectuer les rotations nécessaires.

Si on prétend approfondir dans l'étude du comportement d'autres statistiques il est indispensable d'utiliser un Bootstrap Total, ce qui implique la comparaison de configurations qui correspondent à des espaces factoriels différents. Pour cette raison, la simple considération des résultats bruts du Bootstrap Total peut conduire à des conclusions erronées. Dans notre cas, initialement seule la première dimension de l'analyse paraissait stable, tandis que les autres trois semblaient montrer une certaine instabilité. Après la rotation Procustes, qui réduit dans la mesure du possible les effets des réflexions, échanges entre facteurs et changements d'orientation, on montre que tous les facteurs sont stables. Ce travail ne prétendait pas étudier les autres statistiques.

Références

Álvarez R., Bécue M., Lanero J.J. et Valencia O. (2002). Results stability in textual analysis: Its application to the study of the Spanish investiture speeches (1979-2000). In *Actes des JADT 2002 :*1-12.

Chateau F. et Lebart L. (1996). Assessing sample variability and stability in the visualization techniques related to principal component analysis: Bootstrap and alternative simulation methods. In *Proceedings Computational Statistics. COMPSTAT* : 205-210.

Efron B. et Tibshirani R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

Gifi A. (1990). *Nonlinear multivariate analysis*. John Wiley & Sons Ltd.

Krzanowski W.J. (2001). *Principles of Multivariate Analysis. A User´s Perspective*. Oxford Statistical Science Series.

Lebart L. (1976). The significance of eigenvalues issued from correspondence analysis. In *Proceedings in Computational Statistics*. *COMPSTAT*. Physica Verlag : 38-45.

Lebart L., Morineau A. et Piron M. (1998). *Statistique exploratoire multidimensionnelle*. Dunod.

Lebart L., Salem A. et Bécue M. (1999). *Análisis estadístico de textos*. Milenio. Lleida.

Markus M. (1994). *Bootstrap Confidence Regions in Nonlinear Multivariate Analysis*. DSWO Press.

Meulman J.J. (1984). Correspondence Analysis and Stability. *Research Report 84-01*, Dept of Data Theory, Leiden University.

Michailidis G. et De Leeuw J. (1998). The Gify System of descriptive multivariate analysis. *Statistical Science*, vol. (13) : 307-336.

Milan L. et Whittaker J. (1995). Application of the parametric Bootstrap to models that incorporate a singular value decomposition. *Appl. Statist*, vol. (44/1) : 31-49.

O´Neill M. E. (1978). Aymptotic distributions of the canonical correlations form contingency tables. *Australian Journal of Statistics*, vol. (20/1) : 75-82.

Reiczigel J. (1996). Bootstrap tests in correspondence analysis. *Applied Stochastic Models and Data Analysis*, vol. (12) : 107-117.

Text Retrieval with External Information

Silvano Amato, Emilio Di Meglio, Maria Guerra

Dipartimento di Matematica e Statistica, Università "Federico II"
Via Cintia, Monte Sant'Angelo
80126 – Naples – Italy
{silamato, edimegli, mguerra}@unina.it

Abstract

Aim of this paper is to propose a statistical strategy based on the Partial Least Squares Regression (PLSR) in order to exploit external information in Text Mining processes. We focus on Text Retrieval, a procedure aimed at finding interesting information in large textual collections. In order to exploit external information, often available in collections, we assume a dependence structure between two sets of textual variables. By means of PLSR a latent structure taking into account word similarities and external information is obtained. Projection of documents and queries on the obtained latent structure allows effective document retrieval. The suggested strategy is applied to data which consist of two sets of variables: words of journal article abstracts and their respective keywords. The investigated approach enables more effective documents retrieval, as it takes into account the external information given by the keywords.

Keywords: PLS Regression, text mining, text retrieval, LSI.

1. Introduction

Text Mining can be considered as a process aiming at uncovering interesting information in large, non structured textual *corpora*. Text Retrieval is an important task of Text Mining; aim of this technique is to search for a specific piece of information of a specific topic in large textual repositories according to characteristic aspects of their content.

In text repositories, in addition to documents, we usually find information that is discarded by commonly used Text Mining strategies. This external information, if properly considered, could drastically improve performances. For example, in a text repository containing article abstracts, there is information such as keywords, authors or sources that could be very useful either for retrieval either for other Text Mining tasks. Text Mining tasks, in fact, are based on evaluating similarities among documents; these similarities are based on the meanings carried by the documents, and, external information can improve meanings extraction from documents.

In this paper we deal with document vectors arranged, following the *bag of word* encoding, in a lexical table, $\mathbf{F}^*$; a generic element f^*_{ij} of $\mathbf{F}^*$ is the relative frequency of occurrence of word j in document i. The external information is instead encoded as an indicator matrix $\mathbf{Y}$, that cross-tabulates the same documents of matrix $\mathbf{F}^*$ with the words representing the external information. Matrices $\mathbf{F}^*$ and $\mathbf{Y}$ are shown in figure 1. Usually lexical tables are highly dimensional and sparse, therefore a common step in retrieval problems is dimensionality reduction; this is usually performed by means of Singular Value Decomposition (SVD) based methods applied only to matrix $\mathbf{F}^*$.

Here we suggest a strategy for the analysis of lexical tables that takes into account external information. In this framework, non symmetrical relationships between the two groups of variables

$$\mathbf{F}^* = \begin{array}{c} \text{Abstract words} \\ \left[\begin{array}{ccccc} f^*_{11} & f^*_{12} & \cdots & \cdots & f^*_{1p} \\ f^*_{21} & & & & f^*_{2p} \\ & & f^*_{ij} & & \\ \vdots & & & & \vdots \\ f^*_{n1} & & \cdots & & f^*_{np} \end{array}\right] \end{array} \qquad \mathbf{Y} = \begin{array}{c} \text{Keywords} \\ \left[\begin{array}{cccc} 1 & 0 & \cdots & 1 \\ & & 1 & \\ 0 & & \vdots & 0 \\ \vdots & & 0 & \vdots \\ & \cdots & & \\ 1 & & & 1 \end{array}\right] \end{array}$$

Figure 1. Matrices $\mathbf{F}^*$ *and* $\mathbf{Y}$.

is supposed.

The first group of variables (*independent*), arranged in matrix $\mathbf{F}^*$ is constituted by the words observed in texts; the second group (*dependent*), $\mathbf{Y}$, can be formed by the vocabulary of the external information.

The matrix $\mathbf{F}^*$ is weighted by marginal distributions of rows and columns so to consider chi-square metric. This is done in order to cope with two problems related to word frequency data, namely: different length of documents and triviality of high frequency words (Di Meglio, 2003). In fact, a word that appears only one time in a very short document cannot be considered less important than a word appearing 100 times in a 300 pages book. In most cases, furthermore, rare words are more informative than high frequency ones.

Let $\mathbf{D}_r^{-1}$ and $\mathbf{D}_c^{-1}$ be the diagonal matrices of the marginal row and column distribution, the weighted matrix $\mathbf{F}$ is:

$$\mathbf{F} = \mathbf{D}_r^{-1}\mathbf{F}^*\mathbf{D}_c^{-1} \tag{1}$$

The modeling of the non symmetrical relationship between $\mathbf{F}$ and $\mathbf{Y}$ and dimensionality reduction are jointly achieved by means of Partial Least Squares Regression (PLSR) (Wold *et al.*, 1984) described in section 2.

The PLSR Retrieval strategy, here proposed, performs retrieval in the subspace spanned by the columns of matrix $\mathbf{Y}$, in which the units (documents) of matrix $\mathbf{F}$ are projected.

The subspace generated by the PLSR components can also be effectively used for documents clustering. In this subspace, in fact, similarities among documents are based on semantic similarities and external information. In this work, indeed, we concentrate on Text Retrieval.

2. PLS Regression

PLS Regression has been introduced by Svante Wold *et al.* (1984) in order to model the non symmetrical relationship between two groups of variables with aim of maximizing predictive power of the model and to cope with multicollinearity among variables. From the latter perspective, PLS is an alternative to Principal Components Regression (PCR) (Massy, 1965); literature (e.g. Næs and Martens, 1985; de Jong, 1993) shows that PLS leads to more parsimonious model.

PLS Regression can be interpreted as a penalized Canonical Correlation Analysis (CCA), with a PCA in the $\mathbf{F}$ space and a PCA in the $\mathbf{Y}$ providing the penalties. Namely, Höskuldsson (Höskuldsson, 1988) shows that in the first step, PLS algorithm performs SVD of $\mathbf{F}'\mathbf{Y}$ (if $\mathbf{F}$ is the regressor matrix and $\mathbf{Y}$ is response the one).

Let r be the rank of $\mathbf{F}$; we define $\mathbf{F}_0 = \mathbf{F}$ and $\mathbf{Y}_0 = \mathbf{Y}$. First pair of PLS components $\mathbf{t}_1 = \mathbf{F}_0\mathbf{w}_1$ and $\mathbf{u}_1 = \mathbf{Y}_0\mathbf{c}_1$ are such that

$$\begin{cases} \max\limits_{\mathbf{w}_1,\mathbf{c}_1} \text{cov}(\mathbf{F}_0\mathbf{w}_1, \mathbf{Y}_0\mathbf{c}_1) \\ \mathbf{w}_1'\mathbf{w}_1 = 1 \\ \mathbf{c}_1'\mathbf{c}_1 = 1 \end{cases}$$

where $\text{cov}(\cdot,\cdot)$ stands for covariance. We now denote residual matrix $\mathbf{F}_1 = \mathbf{F}_0 - \mathbf{t}_1\mathbf{p}_1$ and $\mathbf{Y}_1 = \mathbf{Y} - \mathbf{t}_1\mathbf{c}_1$, where $\mathbf{p}_1 = \mathbf{F}_0'\mathbf{t}_1/\mathbf{t}'\mathbf{t}_1$ and $\mathbf{c}_1 = \mathbf{Y}'\mathbf{t}_1/\mathbf{t}'\mathbf{t}_1$. Second pair of PLS components $\mathbf{t}_2 = \mathbf{F}_1\mathbf{w}_2$ and $\mathbf{u}_2 = \mathbf{Y}_1\mathbf{c}_2$ are such that

$$\begin{cases} \max\limits_{\mathbf{w}_2,\mathbf{c}_2} \text{cov}(\mathbf{F}_1\mathbf{w}_2, \mathbf{Y}_1\mathbf{c}_2) \\ \mathbf{w}_2'\mathbf{w}_2 = 1 \\ \mathbf{c}_2'\mathbf{c}_2 = 1 \\ \mathbf{w}_1\mathbf{F}_0'\mathbf{F}_1\mathbf{w}_2 = 0 \end{cases}$$

Last constraint in the above system ensures $\mathbf{t}_1$ and $\mathbf{t}_2$ to be orthogonal to each other. A generic pair h of PLS component is given by solving

$$\begin{cases} \max\limits_{\mathbf{w}_h,\mathbf{c}_h} \text{cov}(\mathbf{F}_{h-1}\mathbf{w}_h, \mathbf{Y}_{h-1}\mathbf{c}_h) \\ \mathbf{w}_h'\mathbf{w}_h = 1 \\ \mathbf{c}_h'\mathbf{c}_h = 1 \\ \mathbf{w}_h'\mathbf{F}_{h-1}'\mathbf{F}_{j-1}\mathbf{w}_j = 0, \forall j < h \end{cases}$$

Columns of matrix $\mathbf{W}_{(h)}$, $w_1, \mathbf{w}_2, \ldots, w_h$, give PLS components $\mathbf{t}$ by means of residual matrix $\mathbf{F}_{h-1}$; in order to compute the $\mathbf{t}$'s by means of original matrix $\mathbf{F}$, we use a transformation of $\mathbf{W}_{(h)}$:

$$\widetilde{\mathbf{W}}_{(h)} = \mathbf{W}_{(h)} \left(\mathbf{P}_{(h)}'\mathbf{W}_{(h)}\right)^{-1}$$

At each step $h = 1, 2, \ldots, r$, PLS regression maximizes covariance between components of residuals matrices $\mathbf{F}_h$ and $\mathbf{Y}_h$; this corresponds to the inter–battery analysis by Tucker (Tenenhaus, 1998; Tucker, 1958) of matrices $\mathbf{F}_h$ and $\mathbf{Y}_h$; that is, because $\text{cov}(\mathbf{F}_{h-1}\mathbf{w}_h, \mathbf{Y}_{h-1}\mathbf{c}_h)$ can be expressed as the product of correlation between $\mathbf{F}_{h-1}\mathbf{w}_h$ and $\mathbf{Y}_{h-1}\mathbf{c}_h$ and $\sqrt{\text{var}(\mathbf{F}_{h-1}\mathbf{w}_h)\text{var}(\mathbf{Y}_{h-1}\mathbf{c}_h)}$. We can see PLS regression as a compromise between canonical correlation analysis (maximum correlation between $\mathbf{t}_h$ and $\mathbf{u}_h$) and OLS regression on principal component analysis (maximum variance of $\mathbf{t}_h$ and $\mathbf{u}_h$).

A fundamental issue in PLS Regression is selection of the number of components to include in the model by means of which we can tune estimates efficiency and distortion; most applied selection method is Cross Validation (e.g. Stone, 1990).

3. Text Retrieval

The classical problem in Text Retrieval is the search for a specific piece of information of a specific topic in large document repositories. In practice, using this methodology, an user should be able to retrieve the relevant documents given a certain natural language query.

A standard Text Retrieval method builds an index of documents and gives the user the possibility to perform searches in this index by formulating queries. Queries are usually formulated in natural language and express the concept the user wishes to retrieve. The system should then be able to compare the concept expressed in the query with all the documents, rank the documents

in order of relevance and give back to the user the n most relevant ones. Text retrieval deals with retrieving contents, but contents can be expressed in many different ways using different words. A retrieval system should therefore be able to extrapolate concepts from documents and assess their similarities with the queries.

Text retrieval strategies

Three of the main strategies for retrieving documents from huge textual databases will now be presented: Boolean Model, Vector Space Model (VSM) and Latent Semantic Indexing (LSI). All these methods, except Boolean Model which is the most simple, are based on the *bag of words* documents encoding.

Boolean search is very similar to a search process in a classical database. In Boolean retrieval the system selects the documents that satisfy a logical expression of query terms using boolean operators such as AND, OR, NOT. This method is usually applied on textual repositories that have already been manually indexed with keywords; in this case queries are boolean expressions of keywords. Boolean model suffers from some serious drawbacks (Salton *et al.*, 1983): e.g., the number of retrieved documents depends on the frequency of the terms used in the query and on the used boolean operators.

Vector Space Model (VSM) is widely used in Text Retrieval mainly because it is easy to implement and is conceptually simple. Proximities among documents are in fact assumed to be similar to proximities in a multidimensional space and this makes possible to use statistical models for building retrieval systems. In VSM each document and each query are represented by vectors of term weights (Salton, 1989). In VSM Text Retrieval is performed by measuring the similarity between the query vector and the document vectors. This similarity can be measured in terms of Euclidean distance or in terms of angle between vectors. The documents that present a larger similarity or a smaller angle to the query are retrieved as relevant. Also, this scheme presents some drawbacks: it does not allow to consider synonymous and then dependencies.

Latent Semantic Indexing (LSI), introduced by Dumais *et al.* (Dumais *et al.*, 1988; Deerwester *et al.*, 1990) is a technique aiming at extracting the hidden semantic structures from texts by means of Singular Value Decomposition (SVD). The technique consists in projecting queries and documents in a space with "latent" semantic dimensions determined using SVD. The latent dimensions derive from the co-occurrence patterns among words. In this way a query and a document can have a high similarity even if they don't share any term (Manning, 2001) as long as the terms are semantically similar according to the LSI analysis. The assumption at the basis of LSI model is that there is an underlying latent semantic structure in word usage that is obscured by noise and by variability of word choice (Dumais *et al.*, 1988). SVD is used to capture the significant information and discard the noise.

LSI is based on Euclidean distance and this distance is considered not adequate for count data contained in lexical tables. Thus document vectors are not well represented in the latent space built by simply calculating SVD of matrix $\mathbf{F}^*$. This procedure, in fact, could be adequate for tables of continuous measurements; in a lexical table, instead it would give the same importance to a term appearing the same number of times in documents of extremely different lengths. In order to cope with these problems it has been proposed to use chi-square metric for text retrieval with LSI (Di Meglio, 2003).

4. Using external information in text retrieval

Text Retrieval is usually performed on the actual texts without considering any other information. A human user, when assessing the similarity between two documents or the relevance of a document uses also external information or metadata to support his/her decision. This information could be given by keywords, abstract, author, source and so on. Textual corpora, in fact, are formed by the actual documents and by other elements that can be explicit: titles, authors, keywords, classifiers, headers, or implicit: collocation, source, time and occasion of issue etc. This information, if properly considered, could drastically improve the performance of a retrieval system. Here we suggest a strategy that aims at exploiting external information available in the data.

PLS Regression is adopted to model the non symmetrical relation between the two groups of variables introduced in section 1. As described in section 2, the first step of PLS performs a SVD of $\mathbf{F}'\mathbf{Y}$. Being $\mathbf{Y}$ an indicator matrix, non zero elements of $\mathbf{F}'\mathbf{Y}$ represent co–presence of keywords and documents terms.

Proposed retrieval strategy is carried out in three steps:

- PLS Regression of $\mathbf{F}$ on $\mathbf{Y}$; computation of PLS components $(\mathbf{T})$ and selection of relevant components by means of the Q^2 index (Tenenhaus, 1998).
- The projection of documents $(\mathbf{t}_i)$ and query $(\mathbf{z}_q)$ vectors on the built subspace, are given by:

$$\mathbf{z}^q = \mathbf{v}^q \widetilde{\mathbf{W}}_{(h)} \tag{2}$$

$$\mathbf{t}^i = \mathbf{f}^i \widetilde{\mathbf{W}}_{(h)} \tag{3}$$

  $\mathbf{z}^q$ ($q = 1, 2, \ldots,$ number of queries), and $\mathbf{t}^i$ ($i = 1, 2, \ldots, n$) are row vectors with h elements, where h is the selected number of components:

- Retrieval of relevant documents to query q, by means of the cosine of the angle between the projected query vector $\mathbf{z}^q$ and all projected documents vectors $\mathbf{t}^i$:

$$d_q^i = \frac{\mathbf{z}^{q\prime}\mathbf{t}^i}{\|\mathbf{z}^q\| \, \|\mathbf{t}^i\|} \tag{4}$$

  where $\|\cdot\|$ is the quadratic norm. For each query q, the distances d_q^i obtained from (4) are sorted in ascending order and, the first k documents are retrieved as the most relevant to query q (i.e. the closest), according to the number of documents required by the user.

5. Application to OHSUMED data

Proposed strategy has been applied to a dataset extracted by the OHSUMED collection (available from ftp://medir.ohsu.edu in the directory /pub/ohsumed) compiled by William Hersh this collections consists of medical journal abstracts. For each abstract author, keywords and journal title are provided.

We randomly selected a subset of 963 documents with respective keywords and used the latter as external information. 10 queries were built and relevance of documents to these queries has been assessed by three independent judges.

PLS has been run and first 36 components, yielding 33.3% of explained inertia, have been retained according to the Stone–Geisser Q^2 index.

The proposed method has been compared with LSI and LSI using chi-square metric (LSCI) on the matrix $\mathbf{X}$, and with LSI on matrix $[\mathbf{X}|\mathbf{Y}]$; the performances of the different methods has been measured by means of the Precision Index (Manning and Schütze, 2001). In order to make fair comparisons among the different methods we retained for each method applied a number of components such that achieve 33.3% of inertia. Table 1 shows the retained numbers of components.

	Components Number
PLS	36
LSI $\mathbf{X}$	81
LSCI	81
LSI $[\mathbf{X}|\mathbf{Y}]$	100

Table 1. Number of components needed to achieve 33.3% of total inertia.

Precision Index measures the proportion of documents that the system got right and ranges between 0 and 1. Precision has been calculated, to take into account the ranking, at different *cutoff* points, that is, considering different numbers of retrieved documents. Usual cutoff points are 5, 10, 20 documents. We have used cutoff 1,2,...,10. In figure 2 precision plot of the four used methods is shown. PLS retrieval globally outperforms the other methods. In fact PLS at cutoffs 1 to 8 performs better or at least equally than other methods. For cutoffs 9 and 10 LSI on $[\mathbf{X}|\mathbf{Y}]$ performs best, but this method performs poorly at lower cutoffs. In any case, PLS ranks the relevant documents always in the first positions.

In particular, after a deeper examination of queries and retrieved documents it appears that LSI and LSCI tend to give higher rank to documents in which the words used in the query have high occurrence. PLS retrieval strategy instead retrieves documents more similar to the query in terms of *meanings*, even if they don't share words with the query. This happens because also keywords are taken into account in the proposed retrieval strategy.

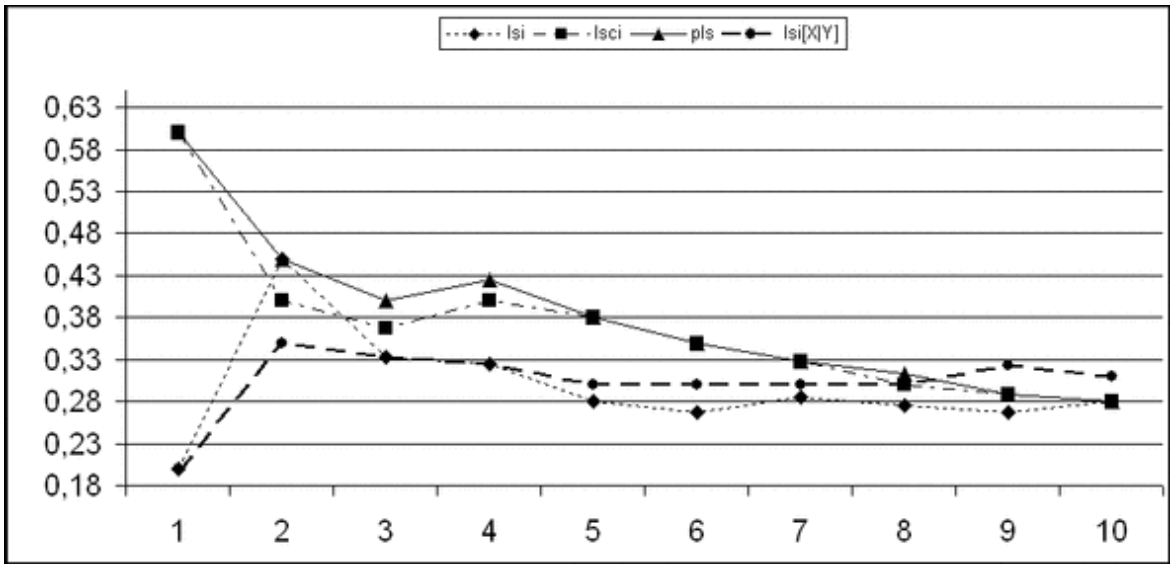


Figure 2. Precision Plot

6. Conclusions and perspectives

On the used data PLS retrieval outperforms both LSI and LSCI. This is due to the use of external information considered in the retrieval strategy.

Massive experiments need however to be performed in order to ascertain pros and cons of the proposed methodology, namely in order to describe the situations in which PLS retrieval is more appropriate than other methods. Further research is needed in order to understand the relationship between the distances computed in the projected documents and their distances in the original space.

Dimensionality reduction obtained by means of PLS can be easily exploited also for clustering problems. Typologies obtained in this manner use in fact external information to classify documents and can lead to more effective cluster identification.

The interpretation of components is also possible; this task can be carried out by means of the projection of words of both matrices $\mathbf{F}$ and $\mathbf{Y}$ on the same subspace. The scatter plot of these projection can be easily read by looking at Euclidean distances in the plane as well as at correlations between words.

Future research directions will include clustering with external information.

Acknowledgment

The authors wish to thanks the unknown referees for their useful suggestions.

References

Balbi S. and Giordano, G. (2000). Un'analisi dei dati testuali con informazioni esterne: le definizioni di "qualità". In *Proceedings of JADT 2000*.

Barker M. and Rayens W. (1999). Partial Least Squares for Discrimination. In *Nonlinear Models Conference* at UK.

Deerwester S., Dumais S., Furnas G., Landauer T. and Harshman R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, vol. (41/6): 391-407.

de Jong S. (1993). PLS fits closer than PCR. *Journal of Chemometrics*, vol. (7): 551-557.

Di Meglio E. (2003). Improving Text Retrieval with Latent Semantic Indexing using Correspondence Analysis. In *Atti del Convegno SIS*.

Dumais S.T., Furnas George W., Landauer Thomas K., Deerwester S. and Harshman R. (1998). Using latent semantic analysis to improve access to textual information. In *Proceedings of the Conference on Human Factors in Computing Systems CHI'88*.

Fisher R.A. (1936). The use of Multiple Measurement in taxonomic problems. *Ann. Eugen.*, vol. (7): 179-188.

Höskuldsson A. (1988). PLS Regression methods. *Journal of Chemometrics*, vol. (2): 211-228.

Manning C.D. and Schütze H. (2001). *Foundations of Statistical Natural Language Processing*. MIT Press.

Massy W.F. (1965). Principal Components Regression in exploratory statistical research. *Journal of the American Statistical Association*, vol. (60): 234-246.

Næs T. and Martens H. (1985). Comparison of prediction methods for multicollinear data. *Communication in Statistics — Simulation and Computation*, vol. (14/3): 545-576.

Salton G., Fox E. and Wu H. (1983). Extended boolean information retrieval. In *Communication of the ACM*, vol. (26/11): 1022-1036.

Salton G. (1989). *Automatic Text Processing*. Addison Wesley.

Stone M. and Brooks R.J. (1990). Continuum Regression: Cross–validated sequencially constructed prediction embarcing Ordinary least Squares, Pratial least Squares and Principal Component Regression. *J. R. Statist. Soc., B*, vol. (52/2): 237-269.

Stone M. (1974). Cross–validatory choice and assesment of statistical prediction. *J. R. Statist. Soc., B*: 237-269.

Tenenhaus M. (1978). *La Régression PLS, Théorie et Pratique*. Éditions Technip.

Tucker L.R. (1958). An inter-battery method of factor analysis. *Psychometrika*, vol. (23/2): 111-136.

Wold S., Rube R., Wold H. and Dunn W. (1984). The collinearity problem in linear regression.The PLS approach to generalized inverses. *SIAM Iorunal of Sci. Stat. Comput.*, vol. (5): 735-743.

Clustering Algorithms for Noun Phrase Coreference Resolution

Roxana Angheluta, Patrick Jeuniaux, Rudradeb Mitra, Marie-Francine Moens

Katholieke Universiteit Leuven – Belgium

{roxana.angheluta, patrick.jeuniaux, marie-france.moens}@law.kuleuven.ac.be,
rudradeb_mitra@rediffmail.com

Abstract

In this paper, we present four clustering algorithms for noun phrase coreference resolution. We developed two novel algorithms for this task, a fuzzy algorithm and its hard variant and evaluated their performance on two different sets of texts in comparison with an existing fuzzy and a hard clustering algorithm that are described in the literature. Our algorithms perform slightly better and do not rely on a predefined threshold distance value for cluster membership. In addition, our fuzzy clustering algorithm seems to perform better than a hard clustering on a pronoun resolution task.

Keywords: Coreference resolution, clustering.

1. Introduction

Most of the natural language processing applications that deal with meaning of discourse imply the completion of some reference resolution activity. The first kind of reference resolution that appears significant in this framework is noun phrase coreference resolution. This is the ability to relate each noun phrase in a text to its referent in the real world. Our coreference resolution focuses on detecting "identity" relationships (i.e. not on is-a or whole/part links for example). Two entities are considered as coreferents when they both refer to the same noun phrase in the situation described by the text (e.g., in the sentences: "Dan Quale met his wife in college. The Indiana senator married her shortly after he finished his studies": "his", "Indiana senator" and "he" all co-refer to "Dan Quale").

It is natural to view coreferencing as a partitioning or clustering of the set of entities. The idea is to gather coreferents into the same cluster, which is accomplished in two steps: 1) detection of the entities and extraction of a specific set of their features; 2) clustering of the entities. For the first subtask we use the same set of features as in Cardie and Wagstaff (1999). We implemented two novel algorithms for the second step: a progressive fuzzy clustering algorithm and its hard variant. We also implemented the hard clustering algorithm presented in Cardie and Wagstaff (1999) and a fuzzy clustering algorithm as described in Bergler *et al.* (2003). Our goal is to test the quality of the coreference resolution that is achieved by these four algorithms.

Coreference resolution is very valuable in text based applications including information retrieval, text summarization, information extraction and question answering, as it refines the representations made of the content of a text.

The next section explains the methods used for feature selection and the four clustering algorithms. We then describe our experiments and their results. In the discussion the four algorithms are critically evaluated.

Feature f	Weight w_f	$function_f$
Words	10.0	(#of mismatching words)/(# of words in longer NP)
Head Noun	1.0	1 if the head noun differs; else 0
Position	5.0	difference in position/maximum difference in document
Pronoun	5.0	1 if NP_i is pronoun and NP_j is not; else 0
Article	5.0	1 if NP_j is indefinite and not appositive; else 0
Words-Substring	-∞	1 if NP_i subsumes NP_j
Appositive	-∞	1 if NP_j is appositive and NP_i is its immediate predecessor; else 0
Number	∞	1 if they do not match in number; else 0
Proper Name	∞	1 if both are proper names but mismatch in every word; else 0
Gender	∞	1 if genders are contradictory; else 0
Semantic class	∞	1 if entities have different semantic classes; else 0

Table 1. Set of features and their weights used in coreference resolution

2. Methods

2.1. *Feature Selection*

An entity is a noun phrase in the text. Each entity is represented as a set of 11 features (see table 1). Here follows a short description of their definition and mode of extraction:

Individual Words: The words of the noun phrase are used to measure the mismatching words between two entities and to see if one entity subsumes (includes totally as a substring) the other.
Head Noun: The noun in the phrase that is not the modifier of another noun.
Position: Each entity is indexed from 1 to n, according to their order of occurrence in the text.
Pronoun: A pronoun is recognized by its part-of-speech (POS) tag.
Article: A noun phrase can be either Indefinite (contains 'a' or 'an'), Definite (contains 'the') or None (without article).
Appositive: An entity is considered appositive if it is enclosed between "," and ",";" or ".", it is a proper name or contains an article and it is immediately preceded by another noun phrase in the text.
Number: The number is detected using the POS tag.
Proper Name: Proper Names are identified by looking at the capitalization of the words.
Gender: A list[1] is used to distinguish the male and female first names. Short lists with words like *brother, mother, etc.* are used to assign gender to common nouns. The rest of the nouns remain unresolved.
Semantic Class: Here we use WordNet (Miller *et al.*, 1990) to classify objects, human entities and pronouns into three different semantic classes: 0, 1 and 2 respectively (S_0, S_1, S_2).
In contrast with Cardie and Wagstaff (1999), we do not include the animacy feature, which indicates whether or not the entity is a living thing.

2.2. *Distance Metric*

The clustering algorithms use the following metric for computing the distance between two entities NP_i and NP_j:

$$dist(NP_i, NP_j) = \sum_{f \in F} w_f * function_f(NP_i, NP_j) \quad (1)$$

where F corresponds to the entity feature set, w_f is the weight of a feature, $function_f$ returns a value in [0,1] (see table 1). A weight of ∞ has priority over $-\infty$: if two entities mismatch on a feature which has a weight of ∞, then they have a distance equal to ∞ (so they are not

[1] http://www.census.gov/genealogy/names.

considered as coreferents). The approach is consistent with Cardie and Wagstaff (1999).

3. The Clustering Methods

3.1. *Hard Clustering Cardie et al. (HC-C)*

In the algorithm of Cardie and Wagstaff (1999) each entity initially forms its own cluster (i.e. a singleton cluster). The algorithm starts from the end of the document and goes backwards, while each noun phrase cluster is compared with all preceding clusters. If the distance (1) between two noun phrases of the two compared clusters is less than a distance value (a threshold set by experiments), their clusters merged, provided all entities are compatible (i.e. don't have a pairwise distance of ∞).

We pick for each cluster a representative member (medoid), chosen as the first person name in the text, if it exists, otherwise as the first entity in the text that appears in the cluster.

The algorithm is very simple and fast, however it has also some weak points.

- The highly greedy character of this algorithm (as it considers the first match and not the best match) introduces errors which are further propagated as the algorithm advances. In the example *"Robert Smith lives with his wife ... Smith loves her"*., when clustering the entities *"Robert Smith"*, *"wife"*, *"Smith"* and *"her"*, the gender of *"Smith"* cannot be determined as it can be both male or female. *"Smith"* may come in the same cluster with *"her"*, if allowed by the distance threshold, but then *"Robert Smith"* will never be correctly resolved with the entity *"Smith"* because of the incompatibility between *"Robert Smith"* and *"her"*.
- The algorithm is very dependent on the threshold distance value for cluster membership. The value of this distance is determined experimentally and may be tuned for each document, which makes the algorithm corpus-dependent.
- The single pass algorithm is very dependent on the order when comparing clusters. Often there are different possibilities for merging clusters - especially when the distance threshold for cluster membership is set to a very high value - , each of them resulting in a different clustering.
- For a high threshold, the algorithm has the tendency to group all entities with semantic class 0 or semantic class 2 in one cluster.

3.2. *Fuzzy Clustering Bergler et al. (FC-B)*

Another promising approach found in the literature (Bergler *et al.*, 2003) considers noun phrase coreference resolution as a fuzzy clustering task because of the ambiguity typically found in natural language and the difficulty of solving the coreferents with absolute certainty. In this algorithm each entity initially forms its own cluster (whose medoid it is). Each other entity is assigned to all of the initial clusters by computing the distance (conform eq. 1 normalized) between it and the medoid of the cluster. Bergler *et al.* (2003) uses additional WordNet (Miller *et al.*, 1990) dependent heuristics for computing the distance metric, which we did not include in our implementation. As a result each entity has a fuzzy membership with each cluster, forming a fuzzy coreference chain (a fuzzy set). Each fuzziness number is in [0,1]. Our implementation of fuzziness is based on the distance function and so it indicates how far we are from coreference. The medoid entity that originally formed the singleton cluster has a complete membership with itself or a distance of zero with itself. Then the chains are iteratively merged when their fuzzy

set intersection is no larger than an a priori defined distance, in other words when there is at least one noun phrase in the fuzzy set of both clusters that has a distance smaller than a threshold with the respective medoids of the clusters. In the merged chain, the medoids of both original chains get complete membership and the membership of the other entities is updated by taking the minimum distance to the medoids. The merging continues until no chains can be merged anymore. The merging assumes that coreference resolution is symmetric and transitive. The algorithm has the effect of a single link hierarchical clustering. Once no more merging can be done, one can form hard clusters by defuzzification. We do that by assigning each entity to the cluster with whom it has the lowest fuzziness.

Beside the fuzzy representation, there are two main differences with Cardie and Wagstaff (1999): 1) the chaining effect is larger because two clusters can be merged even without checking any pairwise incompatibilities of cluster objects; 2) in contrast with Cardie and Wagstaff (1999), the algorithm is independent of the order in which the clusters are merged.

3.3. *Progressive Fuzzy Clustering (FC-P)*

It seemed to us that the previous algorithms exhibit some good points and that a fuzzy clustering is appropriate for the noun phrase coreference task, but the fuzziness could be exploited in such a way that — in contrast with Bergler *et al.* (2003) — uncertainty of cluster membership of all fuzzy members plays a role in cluster merging. In addition, an algorithm that does not rely on a corpus-dependent distance threshold appears more practical.

The algorithm Mitra *et al.* (2003) is summarized in figure 1.

Our clustering algorithm is neither pure hard nor pure fuzzy. Initially, all the entities with semantic class 0 or 1 ($\in S_0$, $\in S_1$) form medoids of singleton clusters. Thus the number of clusters is equal to the number of entities $\in (S_0 \cup S_1)$. For each of the other entities ($\in S_2$, pronouns), a fuzzy membership value is calculated using the distance between the entity and the cluster. The fuzzy set of each cluster is used for merging two clusters. The idea behind the merging is that two clusters that corefer should have quite similar fuzzy sets. The possible noise from two or three entities in the fuzzy set must be smoothed by the rest of them.

To improve the performance two special cases are included in the algorithm.
1) *Appositive merging*: Appositives have much higher preferences than the other features. Thus all the appositives are merged immediately after the formation of the initial clusters.
2) *Restriction on pronoun coreferencing*: According to this restriction, no pronoun can corefer to an element of a cluster whose central medoid is occurring after it. This restriction however prohibits cataphoric references (e.g. *"After he saw the danger, Quayle got scared"*), but they appear quite rarely in texts.

The main resemblances and differences with the foregoing algorithms are:

- *Progressive nature*: As in Bergler *et al.* (2003), our fuzzy algorithm progressively updates the fuzzy membership after each merging of clusters. However it updates it differently, i.e., not by taking the minimum fuzziness of an entity in the merged clusters, but by recomputing the fuzzy membership of an entity in the new cluster (see eq. 2). This is necessary as it is not always possible to correctly resolve some features (e.g. the gender) of a name. Initially the pronouns (he and she) are assigned a fuzzy membership to the clusters. As clusters are merged the feature of the entity may be resolved and thus the fuzzy set membership may change completely.

MAIN

1. Set each entity with semantic class 0 or 1 as the medoid of a singleton cluster.
2. Merge those medoids which have appositive distance -∞.
3. For each entity NP_i with semantic class 2
   For each cluster C_j
   Compute $fuzziness(NP_i, C_j)$ (conform eq. 2).
4. Repeat
   - Find the most similar clusters C_i, C_j.
     - clusters whose medoids have a distance of 0 or $-\infty$ or
     - clusters that have minimum output of $SIMILAR(C_i, C_j)$ and for which $ALL\_NPs\_COMPATIBLE(C_i, C_j)$
   - Merge the most similar clusters.
   - For each NP_i with semantic class 2 (S_2)
     For each cluster C_j
     Recompute $fuzziness(NP_i, C_j)$.
5. Until cannot merge anymore.
6. Defuzzify and assign to each entity of a cluster the coreference represented by the central medoid.

FUZZINESS(NP_i,C_j)

$$fuzziness(NP_i, C_j) = \frac{dist1(NP_i, C_j)}{\sum_k dist1(NP_i, C_k)} \quad (2)$$

where
$dist1(NP_i, C_j) =$
$\sum_{f \in F \setminus position} w_f * function_f(NP_i, NP_k) +$
$+ w_{position} * function_{position}(NP_i, NP_t)$
where NP_k is the central medoid of C_j and, considering the feature text position, NP_t is the closest entity to NP_i which belongs to cluster C_j.

SIMILAR(C_i,C_j)

- $d = 0$
- For each entity NP_k with semantic class 2 for which $fuzziness(NP_k, C_i) < \infty$ and $fuzziness(NP_k, C_j) < \infty$

  $$d = d + fuzziness(NP_k, C_i) - fuzziness(NP_k, C_j) \quad (3)$$

- Return d or ∞ if no NP_k has fuzziness less than ∞ with both C_i and C_j.

MERGE(C_i,C_j)

- Add all medoids from C_j to C_i.
- Set the central medoid of the new C_i as the central medoid of the two merging clusters with the lowest entity index, except when a proper name with a lower index appears in any of the clusters. In the last case, set it to the mentioned proper name.
- Delete C_j.

DEFUZZIFY

- For each entity NP_i with semantic class 2 set $cluster(NP_i) = C_j$ where $C_j = argmin_{C_k}(fuzziness(NP_i, C_k))$.

ALL_NPs_COMPATIBLE(C_i,C_j)

1. For all $NP_a \in C_i$
   For all $NP_b \in C_j$
   If $dist(NP_a, NP_b) = \infty$
   then return FALSE.
2. Return TRUE.

Figure 1. Progressive Fuzzy Clustering Algorithm

- *Merging of clusters*: In contrast to Bergler *et al.* (2003) and Cardie and Wagstaff (1999), our criteria for merging clusters are different by restricting the merging of chains that have a non-pronoun phrase as medoid and by considering the similarity of the current fuzzy sets of the clusters.
- Conform to Cardie and Wagstaff (1999) but unlike Bergler *et al.* (2003), our algorithm does not merge clusters when members of the new cluster would be incompatible (except when their central medoids have a distance of 0 or $-\infty$).
- *Search for the best match*: Conform to Bergler *et al.* (2003), but unlike Cardie and Wagstaff (1999), the algorithm iteratively searches for the best match instead of the nearest match for merging clusters. Thus for the first example given in section 3.1: *"Robert Smith"* would be resolved to *"Smith"* and *"her"* would never be integrated with *"Smith"*.
- *Corpus-independent*: Unlike Cardie and Wagstaff (1999) and Bergler *et al.* (2003), the algorithm is corpus-independent as no threshold distance is used.

3.4. *The Hard Variant (HC-V)*

We also implemented the hard variant of the above progressive fuzzy clustering algorithm. The algorithm is summarized in figure 2.

MAIN

1. Set each entity with semantic class 0 or 1 as the medoid of a singleton cluster.
2. Merge those medoids which have appositive distance -∞.
3. Repeat
   - Find the most similar clusters C_i, C_j.
     - clusters whose medoids have a distance of 0 or $-\infty$ or
     - clusters that have minimum output of $SIMILAR(C_i, C_j)$ and for which $ALL\_NPs\_COMPATIBLE(C_i, C_j)$.
   - Merge the most similar clusters.
4. Until cannot merge anymore.
5. For each NP_i with semantic class 2 (S_2)
   For each cluster C_j
   Compute $fuziness^a(NP_i, C_j)$ conf. eq. 2
   $cluster(NP_i) = C_j$ where
   $C_j = argmin_{C_k}(fuziness(NP_i, C_k))$
6. Assign to each entity of a cluster the coreference represented by the central medoid.

SIMILAR(C_i,C_j)

- $d = 0$, $count = 0$
- For each entity NP_k of C_i
  For each entity NP_l of C_j
  $d = d + dist(NP_k, NP_l)$ (conform eq. 3)
  $count = count + 1$
- Return $d/count$ (or ∞ if $d = \infty$)

[a] The name *fuziness* is used here for the sake of consistency with the previous apparition of the same formula and does not imply any reference to a fuzzy character of the algorithm.

Figure 2. Hard Clustering Variant

We assign the entities of semantic class 2 (i.e. the pronouns) to their closest cluster at the completion of the merging process and not at its start, because we hope to be able to assign them more correctly in case of missing features (e.g., gender) of some of the entities. The difference of this hard algorithm with the foregoing fuzzy algorithm lies in the computation of the similarity between clusters when merging the most similar clusters. In our fuzzy algorithm we use the complete fuzzy set of the clusters in the computation of the similarity (eq. 2); in the hard algorithm we merge clusters with a group average hierarchical clustering scheme until no more clusters can be merged based on incompatibility of members. The difference with the fuzzy clustering of Bergler *et al.* (2003) lies in the merging of entities that only belong to S_2 and the use of a group average hierarchical clustering scheme instead of a single linkage hierarchical clustering scheme in Bergler *et al.* (2003).

Preliminary tests showed that pronouns like *it*, *I*, *me*, *our* are very difficult to be resolved and they harm the performance. So in all the above clustering algorithms we considered them alone in singleton clusters, which is not a correct solution, but the results will be affected equally for all the algorithms.

4. Corpora and Evaluation

We used two corpora: one from the Document Understanding Conference 2002[2], the other from the Message Understanding Conference 6[3]. The DUC documents were selected from the category "biographies" and they are small texts (on average 3KB each) which contain many entities to be resolved (pronominal and non pronominal entities). We chose randomly ten documents

[2] Document Understanding Conference http://www-nlpir.nist.gov/projects/duc/guidelines/2002.html.

[3] Message Understanding Conference http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html.

from this set, parsed them in order to extract the entities (as the smallest noun phrases which do not contain embedded noun phrases) and annotated them manually for coreference. The MUC-6 proposed a training set of 30 documents (the so-called "dryrun set") and a test set of 30 documents (the "formaltest set"). They are all annotated with coreference information. In this corpus, both the smallest noun phrases and the ones which contain them are considered as entities. The features are extracted slightly differently for the two corpora, because of the different nature of the entities (e.g. the proper name feature is reduced to the capitalization of the head noun in the MUC corpora, while in the DUC subcorpus we look at two consecutive capitalized words).

The MUC-6 corpora contains few pronouns, which are the entities that are the most interesting for our algorithm. That's why the DUC subcorpus is useful for the evaluation, especially for the pronoun resolution. The DUC documents have an average of 18 pronouns comprised in the set *"he", "she", "him", "her", "they", "them"*, while the two MUC-6 corpora have only a mean of 2.97 and 4.86 respectively.

We computed automatically the precision and recall and combined them into the F-measure. Two algorithms were initially implemented to perform the evaluation: the one of Vilain *et al.* (referred in Baldwin *et al.* (1998), employed in MUC) and the B-CUBED algorithm, described in Baldwin *et al.* (1998).

In Vilain's algorithm, the recall is computed as follows:

$$R = \frac{\sum_i(|C_i| - |p(C_i)|)}{\sum(|C_i| - 1)}$$

where C_i is a manual cluster and $p(C_i)$ is a partition of C_i relative to the automatic clusters ($|p(C_i)|$ measures how many of the automatic clusters contain entities from the C_i). The precision is computed by inverting the roles of the manual and automatic clusters.

In the BCUBED algorithm, the recall is computed as follows:

$$R = \frac{\sum_i R_i}{n}$$

where R_i is the recall obtained for entity i and n is the total number of entities. The recall for entity i is defined as:

$$R_i = \frac{|HC_i \cap CC_i|}{|HC_i|}$$

where HC_i is the manual cluster in which entity i appears and CC_i is the automatic cluster in which entity i appears. The precision is computed by inverting the roles of the manual and automatic clusters.

F-measure combines equally the precision with the recall: $F = \frac{2*P*R}{P+R}$

As observed by Baldwin *et al.* (1998), the algorithm of Vilain yields unintuitive results for some cases, because it does not give any credit for separating out singletons and it considers all types of errors as equal. The algorithm favors big clusters, which is confirmed by the results of our second baseline, which clusters all entities in one cluster (see below) (Vilain F-measure is 86% for the dryrun set, 87% for the formaltest set and 60% for the DUC subcorpus). Because of this, we decided not to use the results obtained with the Vilain's algorithm and we considered only the BCUBED algorithm.

We separately evaluated pronoun coreference, by selecting as entities only pronouns and their immediate antecedents in the manual files. For certain text processing tasks, the correct resolution of pronouns is important. With the Vilain and BCUBED evaluation it is still possible to obtain reasonably good results when a set of pronouns are clustered together, while their antecedent is missing or is wrongly assigned in the cluster. So we computed also the accuracy of the antecedent resolution.

$$accuracy = \frac{\text{number of pronouns correctly resolved}}{\text{total number of pronouns}}$$

A pronoun is considered correctly resolved when its main coreferent in the automatic cluster (the medoid of the cluster) is in the same manual cluster as the pronoun itself and when this coreferent is not another pronoun (case appearing in HC-C and FC-B).

5. Results and Discussion

We tested the algorithms with the two corpora. We tried two baselines : every entity in a singleton cluster (BL1) and all entities in one cluster (BL2). For the hard clustering we used four different threshold values, determined experimentally: 8, 11.5, 16 and 20 (a small threshold corresponds to a conservative behavior: small clusters). For the fuzzy clustering, we used a threshold value of 0.2 and 0.5 (again, a small threshold is conservative).

5.1. *All entities*

The results using all the entities are summarized in table 2.

a)

Algorithm	Precision	Recall	F-measure
BL1	1.00	0.25	0.40
BL2	0.13	1.00	0.22
FC-P	0.81	0.52	0.62
HC-V	0.85	0.51	0.63
HC-C thrs.8	0.83	0.51	0.62
HC-C thrs 11.5	0.57	0.58	0.56
HC-C thrs. 16	0.52	0.61	0.55
HC-C thrs. 20	0.49	0.62	0.54
FC-B thrs. 0.2	0.83	0.51	0.62
FC-B thrs. 0.5	0.39	0.66	0.47

b)

Algorithm	Precision	Recall	F-measure
BL1	1.00	0.24	0.38
BL2	0.19	1.00	0.30
FC-P	0.78	0.48	0.58
HC-V	0.82	0.45	0.57
HC-C thrs.8	0.85	0.43	0.56
HC-C thrs 11.5	0.67	0.52	0.57
HC-C thrs. 16	0.60	0.54	0.55
HC-C. thrs. 20	0.58	0.56	0.56
FC-B thrs. 0.2	0.79	0.46	0.56
FC-B thrs. 0.5	0.40	0.65	0.46

c)

Algorithm	Precision	Recall	F-measure
BL1	1.00	0.56	0.72
BL2	0.07	1.00	0.13
FC-P	0.76	0.77	0.76
HC-V	0.85	0.72	0.78
HC-C thrs.8	0.79	0.67	0.72
HC-C thrs 11.5	0.45	0.73	0.55
HC-C thrs. 16	0.42	0.74	0.53
HC-C. thrs. 20	0.39	0.78	0.52
FC-B thrs. 0.2	0.88	0.67	0.76
FC-B thrs. 0.5	0.26	0.78	0.38

Table 2. Precision, recall and F-measure obtained considering all entities, using the different algorithms for a) dryrun subcorpus of the MUC-6 corpora, b) formaltest subcorpus of the MUC-6 corpora and c) DUC subcorpus.

Here follow a few remarks about the results:

1) The progressive fuzzy algorithm and its hard variant are among the best in terms of F-measure on all corpora.

2) For the HC-C and FC-B, the general tendency is to obtain better results for lower values of the threshold. A more conservative algorithm induces a higher precision, influencing positively the F-measure (although the recall decreases).

3) The difficulty to set an appropriate threshold can be observed in the MUC-6 corpora, since in the training corpus (dryrun), the best results were not obtained with the same threshold as in the test corpus (formaltest).

4) In contrast with precision, which is generally good, the recall values are quite low. We identified three types of errors responsible for the values and we analyzed more deeply the first one (see subsection *Experiment*).

- Wrong assignment of the semantic class. Pronouns (e.g. "he") come in the same clusters with entities like *Bentonville* or *Institute*, which are wrongly considered as person names. This type of errors might be resolved using a name entity recognition tool, able to semantically classify the entities.
- Acronym resolution. The algorithms are not able to relate acronyms with their corresponding long form (e.g. *"International Business Machines Corp." and "IBM"*). This problem is more frequent in MUC corpora (1.8% of the entities in the formaltest subcorpus are acronyms which can not be detected with Word-Substring feature or appositive - like in the case *"the Congressional Black Caucus (CBC)"*). We need to integrate an acronym resolution tool in order to correct this type of errors.
- Discourse structure. The texts contain a lot of direct speech, where the pronouns are very difficult to be resolved. For example, in the following phrase *"He spends most of his time talking to associates and customers," Shinkle said, "and he always comes back with many ideas from them"* (DUC subcorpus), the pronouns *"he"* and *"his"* are wrongly resolved as coreferring with *Shinkle*. This problem is more frequent in the DUC subcorpus (6.52% of the sentences are similar to the above example). A number of discourse specific heuristics could be added to the algorithm.
- Other errors. A number of other errors are due to different causes: lack of synonyms/ hypernyms detection (*"Coca-Cola", "Coke"*, *"million", "cents"*), lack of knowledge of the world (*"the U.S.", "the country"*). These errors are more difficult to resolve.

5.1.1. *Experiment*

In order to quantify errors caused by the wrong assignment of the semantic class, we manually corrected this field and rerun the experiments. The new results are in table 3. The recall increases in all cases. The precision usually decreases for HC-C, which favors big clusters of entities with semantic class 0 (the number of such entities increases by correcting the semantic class), decreasing also the F-measure. For the other algorithms, F-measure usually increases. The influence on the results is larger for the MUC corpora than for the DUC subcorpus, because the number of entities with semantic class wrongly assigned was higher (20.98% for the dryrun and 19.29% for the formaltest comparing with 8.42% for the DUC subcorpus).

5.2. *Pronouns*

Evaluating just the pronouns, the results are summarized in table 4.

Here follow a few remarks about the pronoun resolution:

1) The results obtained for the DUC subcorpus are more representative than the ones obtained

a)

Algorithm	Precision	Recall	F-measure
BL1	1.00	0.25	0.40
BL2	0.13	1.00	0.22
FC-P	0.77	0.55	0.62
HC-V	0.81	0.54	0.64
HC-C thrs. 8	0.81	0.56	0.65
HC-C thrs. 11.5	0.45	0.67	0.53
HC-C thrs. 16	0.40	0.70	0.50
HC-C thrs. 20	0.38	0.73	0.49
FC-B thrs. 0.2	0.79	0.55	0.63
FC-B thrs. 0.5	0.35	0.76	0.47

b)

Algorithm	Precision	Recall	F-measure
BL1	1.00	0.24	0.38
BL2	0.19	1.00	0.30
FC-P	0.78	0.55	0.62
HC-V	0.83	0.50	0.61
HC-C thrs. 8	0.83	0.47	0.59
HC-C thrs. 11.5	0.56	0.59	0.56
HC-C thrs. 16	0.50	0.61	0.54
HC-C. thrs. 20	0.49	0.65	0.54
FC-B thrs. 0.2	0.81	0.48	0.59
FC-B thrs. 0.5	0.42	0.72	0.51

c)

Algorithm	Precision	Recall	F-measure
BL1	1.00	0.56	0.72
BL2	0.07	1.00	0.13
FC-P	0.76	0.80	0.78
HC-V	0.86	0.72	0.78
HC-C thrs. 8	0.78	0.67	0.72
HC-C thrs. 11.5	0.38	0.76	0.51
HC-C thrs. 16	0.36	0.75	0.48
HC-C. thrs. 20	0.33	0.80	0.46
FC-B thrs. 0.2	0.88	0.67	0.76
FC-B thrs. 0.5	0.27	0.78	0.40

Table 3. Precision, recall and F-measure obtained considering all entities, using the different algorithms, starting with correct semantic classes for a) dryrun subcorpus of the MUC-6 corpora, b) formal-test subcorpus of the MUC-6 corpora and c) DUC subcorpus.

for the MUC-6 corpora, since the former contains more pronouns.

2) HC-C and FC-B work now better for higher values of the threshold. This can be explained by the fact that in all corpora most of the pronouns referred to the same person and so they should come out in the same cluster, which rewards big clusters.

3) Since in FC-P we give special attention to the pronouns (by iteratively recomputing the fuzzy vectors), we assume that this algorithm should outperform its hard variant on pronoun resolution. This assumption seems to be verified considering the F-measure, especially in the DUC subcorpus, which is the most representative for this evaluation, but is somehow contradicted by the accuracy results. We need additional tests to sustain our claim.

4) In the DUC subcorpus: although the F-measure is higher for HC-C with threshold 20 than for FC-P, the accuracy measure is lower. We believe that accuracy is a more fair measure for pronoun resolution, since correctly resolving the antecedent of the pronoun is more important than correctly grouping pronouns in the same cluster.

6. Conclusion and Future Work

In this paper we compared four clustering methods for coreference resolution: one progressive fuzzy, its hard variant and another two algorithms, taken from the literature. We evaluated them on two kinds of corpora, a standard one used in the coreference resolution task and another one containing more pronominal entities. Our algorithms are on top when all the entities are considered. For the pronoun resolution, our fuzzy algorithm obtains competitive or better F-measure results compared with the two algorithms from the literature and outperforms both of them in term of accuracy. These results are obtained despite the fact that our algorithms do not rely on a threshold distance value for cluster membership, which makes them corpus-independent. In the future we plan to perform more experiments with different types of texts and to enlarge the feature set based on current linguistic theories. We also plan to integrate the noun phrase coreference tool in our text summarization system.

a)

Algorithm	Precision	Recall	F-measure	Acc
BL1	1.00	0.35	0.51	0.00
BL2	0.39	1.00	0.51	0.32
FC-P	0.93	0.47	0.62	0.09
HC-V	0.94	0.47	0.61	0.14
HC-C thrs. 8	0.94	0.40	0.55	0.00
HC-C thrs. 11.5	0.87	0.41	0.56	0.00
HC-C thrs. 16	0.87	0.41	0.56	0.00
HC-C. thrs. 20	0.84	0.46	0.58	0.09
FC-B thrs. 0.2	0.93	0.41	0.56	0.00
FC-B thrs. 0.5	0.81	0.43	0.55	0.00

b)

Algorithm	Precision	Recall	F-measure	Acc
BL1	1.00	0.27	0.41	0.00
BL2	0.51	1.00	0.62	0.36
FC-P	0.88	0.43	0.55	0.17
HC-V	0.90	0.42	0.54	0.20
HC-C thrs. 8	0.93	0.34	0.48	0.00
HC-C thrs. 11.5	0.89	0.38	0.51	0.01
HC-C thrs. 16	0.88	0.39	0.52	0.01
HC-C. thrs. 20	0.86	0.47	0.58	0.13
FC-B thrs. 0.2	0.91	0.34	0.48	0.00
FC-B thrs. 0.5	0.77	0.40	0.50	0.00

c)

Algorithm	Precision	Recall	F-measure	Acc
BL1	1.00	0.18	0.31	0.00
BL2	0.46	1.00	0.62	0.46
FC-P	0.87	0.55	0.67	0.41
HC-V	0.92	0.45	0.59	0.42
HC-C thrs. 8	0.91	0.32	0.47	0.00
HC-C thrs. 11.5	0.87	0.48	0.62	0.00
HC-C thrs. 16	0.86	0.51	0.63	0.06
HC-C thrs. 20	0.82	0.62	0.70	0.22
FC-B thrs. 0.2	0.91	0.32	0.47	0.00
FC-B thrs. 0.5	0.78	0.53	0.62	0.06

Table 4. Precision, recall and F-measure obtained considering just the pronouns, using the different algorithms for a) dryrun subcorpus of the MUC-6 corpora, b) formaltest subcorpus of the MUC-6 corpora and c) DUC subcorpus.

References

Baldwin B. *et al.* (1998). Description of the UPENN CAMP System as Used for Coreference. In *P*roceedings of the Seventh Message Understanding Conference (MUC-7). http://www.icl.pku.edu.cn/bswen/nlp/www.muc.saic.com/muc_7_toc.html

Bergler S. *et al*. (2003). Using Knowledge-poor Coreference Resolution for Text Summarization. In *Proceedings of Document Understanding Conference 2003*. NIST: 85-92.

Cardie C. and Wagstaff K. (1999). Noun Phrase Coreference as Clustering. In *Proceedings of the Joint Conference on Empirical Methods in NLP and Very Large Corpora*. http://citeseer.nj.nec.com/article/cardie99noun.html

Charniak E. (1999). A Maximum-Entropy Inspired Parser. *Technical Report CS-99-12*, Brown University, August. http://citeseer.nj.nec.com/charniak99maximumentropyinspired.html

Document Understanding Conference http://www-nlpir.nist.gov/projects/duc/guidelines/2002.html

http://www.census.gov/genealogy/names

Miller G.A. *et al.* (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography (special issue)*, vol. (3/4): 235-312.

Message Understanding Conference http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html

Mitra R. *et al.* (2003). Progressive fuzzy clustering for noun phrase coreference resolution. In *Proceedings of the 4'th Dutch-Belgian Information Retrieval Workshop*: 19-30.

Quantification of Stylistic Traits: A Statistical Approach

Mappillairaju Bagavandas, G. Manimannan

Department of Statistics – Madras Christian College – Chennai-600 059 – India
mbdas49@hotmail.com, manimannang@yahoo.co.in

Abstract

It is often recognized that authors have writing styles and it is possible to find a simple statistical mo del, which explains reasonably what makes an author unique. This paper makes an attempt to identify the distinct stylistic features of three Tamil Scholars of the same period and also tries to quantify the writing styles of these authors using eighteen stylistic features. These stylistic features of this study are eleven morphological variables, four habitual words and three function words. ANOVA technique, two sample t-statistic and Factor analysis are used for measuring such stylistics traits and also identifies those traits, which are most densely packed.

Keywords: author style statistics, ANOVA, two-sample t-statistic, factor analysis.

1. Introduction

It has been recognized that an author has a unique writing style which is expressed in the form of subconscious stylistic features. Style of an author can be quantified by counting his\her choice of words for expressing his\ her ideas under the assumption that the writer favouring a stock of words for the expression of ideas is regarded, to some extent, subject to chance (Holmes and Forsyth, 1995). Hence given a certain personality and thus a certain style, as its expression, the characteristic properties of style can be described in terms of statistical law (Herdan, 1964). Bailey (1979) says that the stylistic features of a matured writer will be salient, structural, frequent and easily quantifiable. Thus style reflects personality of a writer and this unconscious process is consistent in the case of matured writers (Holmes, 1985).

Statistical stylistic study not only compliments the traditional scholarship of literary experts but also provides an alternative method for investigating the works of doubtful provenance (Holmes, 1998). These studies provide authentic results if they work within the same genre and also work within as close a time period as possible. Stylistic markers which occur most frequently in a given passage are also identified by these methods (Mealand, 1997). These stylistic studies inhabit two types of problems, the first being the selection of suitable set of stylistic variables and the second being the selection of appropriate techniques. There is no general agreement on the stylistic variable that should be used in stylistic studies. In general, when choosing the stylistic variables, one must use something that has large variation across authors and relatively little variation among an author's own work. Initially, lexical variables have predominated in the stylometry studies, yet this decade has seen the application of syntactic and semantic variables (Holmes, 1998).

Mathematician like Fucks (1952) may be considered pioneers in laying a foundation for more vigorous and objective stylistic analysis through his attempts to quantify stylistic features. Mosteller and Wallace's study (1964) is considered as the first authentic stylometric study

soundly based on modern statistical procedure using computer as its major research tool. John Burrows (1987) through his series of seminal papers introduced stylometry studies as a viable tool for authorship attribution problem. The availability of modern computing facility has provided a unique opportunity for many stylometricians to introduce many multivariate meth ods like factor analysis, cluster analysis and correspondence analysis for conducting experiments with high dimensional data and also to widen the frontiers of stylometry (Peng, 2001).

Factor analysis is considered as an ideal method for determining the relationship between stylistic features and stable personality traits (Sommers, 1966; Herdan, 1964). This analysis helps to find out whether different writers really represent different distinct forms of behaviours or whet her they draw from a limited stock of vocabulary (Miles and Selvin, 1966). This multivariate technique is also used for measuring the extent to which groups of words have similar patterns of high or low use of various writers.

Herdan (1941) was the first to use the factor analysis for analysing the relation between six authors and for identifying the one who uses the most difficult words. This analysis was also applied to establish the common ancestors of a number of proto Indo-European languages (Johnson and Kotz, 1967). Roger Peng and Nicolas Hengartner (2002) have used factor analysis to examine each individual author's function-word counts and also to filter out words which account for very little of the variation between authors in the group. This analysis was used by David Mealand (1997) to establish that samples of different genres from the Gospel of Mark vary in style and also to identify the stylistic markers which are most heavily used in these passages.

2. Data and methods

The present study deals with the literary works of three contemporary Tamil scholars,namely, MahakaviBarathi (MB), V.Kalyanasundaram (VK) and Subramaniya Iyer (SI). In the Pre–Independence period, these three scholars have written number of articles on India's Freedom Movement in the magazine called *India*. Initially, all the three scholars have written articles by attributing their names. The oppressive attitude of the then British Regime made all the three writers to write articles on the same topic anonymously in the same magazine. All the attributed and unattributed articles written on India’s Freedom Movement in that magazine were complied and brought out as a book entitled *Bharathi Dharisanam* in the year 1975. For this quantitative stylistic study, all attributed articles of these three scholars written on India’s Freedom Movement in the year 1906 are considered. Our study is based on nineteen articles of Bharati, six of Kalyanasundram and seven of Subramaniya Iyer.

In stylometry, there are important decisions to be made about the features to be selected and the methods to be used (Mealand, 1997). Eighteen stylistic features are consider ed for this study. They include eleven morphological variables, four habitual words and three function words. The exact lists of variables of this study with their abbreviations are given in Table 1.

For a comparative analysis the frequency counts of the stylistic features must be normalized to the text length in an article. In this study since each sentence is considered as a sample, to normalize the stylistic features, the raw frequency counts of each stylistic feature is divided by the number of words in each sentence and then multiplied by hundred to express it in percentage. Eighteen stylistic features are identified from each sentence. These features include parts of speech, habitual words and function words. Both voices and tenses are expressed in frequencies but not in percentages. If we have n sentences and if we identify p stylistic features

from each sentence, then we have a data matrix of size n x p. Thus, each article was converted as a data matrix and these data matrices form the basis for this quantitative study.

A chi-square analysis of the nineteen articles of Bhararthi establishes that these articles do not differ from one another in terms of the frequency distribution of occurrence of these stylistic features. Similar results were obtained in the case of other two scholars (Manimannan and Bagavandas, 2001). Hence all the nineteen articles of Bharathi are considered as one article for this study. So also, the six articles of Kalyanasundram and seven articles of Subramaniya Iyer. In this study, each sentence is considered as a sample. Hence the nineteen articles of Bharathi consist of three hundred and fifty three sentences, six articles of Kalyanasundram consist three hundred and eighty two sentences and seven articles of Subramaniya Iyer consist of three hundred and fifteen sentences. As there are three authors, there are three data matrices and their sizes are (353x18), (382x18) and (315x18) respectively. Hence the aim is to compare the data matrices of the linguistic features of the three scholars. Average values, two-sample t-statistic values and Euclidean distance values are given in Table1.

3. Analysis

This analysis section consists of two parts. Part one identifies the special stylistics features of each author and Part two quantifies the writing style of each author.

3.1. *Identification of Special Stylistic Features*

This univariate analysis compares the average values of the stylistic features of the three scholars. This comparative study is made in two stages. In the first stage, the hypothesis of equality of the means of a particular feature of three authors is tested using ANOVA (one-way) technique. The acceptance of this hypothesis indicates that particular stylistic feature has no discriminatory power. However, if this hypothesis is rejected, then mean difference of a feature between any two authors is tested using the conventional two-sample t-statistic.

This two-stage comparative analysis indicates that the stylistic features like two-letter word, three-letter word and pronoun do not discriminate these three scholars from one another. That is, these three scholars had the habit of using the same number of these features in writing a sentence. This result indicates that all these three scholars had used, on an average, one pronoun, one two-letter word and two three-letter words in a sentence of ten words. Also the smaller percentages of occurrence of features like intensifier, infinity and adverb indicates that these three authors had used these three features very rarely.

The percentages of occurrences of stylistic features like noun, post-position, clitic, case makers and conjunctions differentiate these three authors statistically from one another. This result indicates that Bharathi is identified as the least user of these features whereas Kalyanasundram is identified as the maximum user of the same stylistic features. Subramaniya Iyer is not identified with any distinct stylistic features because the percentages of the occurrences of stylistic features of this author indicate that the writing style of this author shares equally the special features of the other two authors. The Euclidean distance values confirm this result in Table 1.

This analysis shows that in the sentence of ten words, Bharathi had used, on an average, one postposition, one clitic but three nouns and four case markers and two conjunctions. But on the other hand, in the sentence of the same length, on an average, Kalyanasundram had used five nouns, four conjunctions, three postpositions, three clitics but six case markers. Also it can be seen that the third author, Subramania Iyar had used, on an average, four nouns, three post positions, three clitics, six case markers and four conjunctions.

3.2. *Stylo – Statistical Analysis*

Factor analysis is a variable-oriented multivariate technique. This analysis describes the inter-relationship among many variables in terms of a few underlying, but observable, random qualities called factors (Lawley and Maxwell, 1971). Factor analysis can be considered as an extension of principal component analysis and is used for data reduction and interpretation. This analysis is also used for grouping of variables in such a way that the variables are highly correlated with in groups but have relatively insignificant correlation with variables of different groups. Correlation matrix of the eighteen stylistic features is calculated for each data matrix. The initial statistics are given in Table 2 and groups of stylistic features are given in Table 3.

3.2.1. *The case of Bharathi*

All the eighteen features are highly loaded in the first seven factors, which covers nearly 54 % of the total variation present in this data set. In other words these eighteen features are grouped into seven clusters on the basis of the inter-relationship among themselves. The features like words starting with vowel, verb, two-letter, three-letter and four-letter words are highly loaded in the first factor and hence they form as a cluster. This result shows that the writer Bharathi had preferred verbs and words starting with vowels either as two-letter or three-letter or four-letter words. Since four out of these five features are habitual words, this factor is named as habitual-word factor.

Factor two is highly correlated with features like clitics and case makers. These correlated relationships establish that this writer had the habit of using clitics and case makers in the ratio of 1: 4 in a sentence of ten words. As these two features are function words, this factor is known as function-word factor. Third factor is a contrast between features like noun and pronoun and also they occur in the ratio 4:1 and whenever the occurrence of noun increases the occurrence of pronoun decreases in a sentence. These two features are morphological variables and hence this factor is known as morphological factor. Statistical features like tenses and numeral are accommodated in the fourth factor-the tense factor. This factor indicates that this writer had used to write sentences mostly in past tense with a very few numerals.

Since features like voice and postposition are accommodated in the fifth factor, this factor may be named as voice factor. This factor is a contrast between voice and postposition. This indicates that the writer had favoured to write sentences in the past tense with less number of postpositions. The sixth factor is a syllable factor and it's established that the length of ten words sentence on an average fifteen syllables.

The Seventh factor is contrast between two groups of features. Infinity and adverb are grouped together and intensifiers and conjunction are grouped together. The occurrence of these stylistic features like intensifier, infinity and adverb are rare phenomena. But Bharathi had used at least two conjunctions in a sentence of ten words and hence this factor may be called conjunction factor.

Summarizing, the writer Bharathi had used passive voice sentences in past tense to narrate India's Freedom Movement. In sentence of ten words, he had used, on the average, one clitic, one pronoun, two verbs, three words starting with vowels, four case makers and four nouns. The verbs and words starting with vowels are either two-letter or three-letter or four-letter words. The incre ase in the occurrence of nouns reduces the occurrence of pronouns.

3.2.2. *The case of Kalyanasundram*

The first four factors, which cover nearly 36 % of the total variation present in the data set, had grouped all eighteen features into four clusters. In the first factor features like case maker and clitic are highly accommodated in the ratio 3:1 and hence this factor is known as function-word factor. The features like verb and noun are grouped in the second factor in the ratio 1: 2. This factor is a contrast between verb and noun, which indicates that whenever the occurrence of nouns increases, the occurrence of verbs decreases. This is a morphological factor.

The third factor is a habitual-word factor as it accommodates all the three habitual words with pronoun. This factor is a contrast between four-letter word and the set of two-letter and three-letter words and pronouns. More the occurrence of pronouns, less it will be four-letter words. Fourth factor is a contrast factor. Adverb, syllable, conjunction and infinity are grouped in one set and voice, tense, word starting with vowel, intensifier, and clitics are grouped in another set. This result shows that this author had written active voice sentence in past tense. This author has provided four conjunctions, three words starting with vowels and three clitics. The occurrence of more conjunctions in a sentence reduces the occurrence of clitics and words starting with vowels.

Finally, the author Kalyanasundram used active voice sentences in past tense to describe India's Freedom Movement. In these sentences of ten words, on the average, three postpositions, three clitics, three words starting with vowels, four conjunctions, four nouns and six case markers are accommodated. The occurrence of more verbs red uces the occurrence of nouns; also the occurrence of more conjunctions reduces the occurrences of clitics and the words starting with vowels.

3.2.3. *The case of Subramaniya Iyar*

All the eighteen features are accommodated in the first seven factors, which covers nearly 56 % of variation present in the given data set. In the first factor, verb, word starting with vowel, syllable and four-letter word are accommodated. This indicates that verb and word starting with vowel will be four-letter words with two or three syllables. This is a morphological factor. Case marker and clitic are highly loaded in the second factor and they occur in the ratio 2:1 in a sentence. This is a function- word factor.

Third factor, a contrast factor, provides high loading for nouns and pronouns in the ratio 6:1. The occurrence of more nouns reduces the occurrence of pronouns. This is a noun-family factor. Fourth factor accommodates voice and conjunction. This author used to write passive voice sentences with at least three conjunctions. In the fourth factor adverb and two-letter word are accommodated and this indicates that the adverbs of this author are identified as two-letter words.

Fifth factor is a tense factor. This writer used to write sentence in present tense. The last factor contrasts between two sets of features. In one set postposition, intensifier and three-letter word are accommodated and in the other set numeral and infinity are accommodated. This result indicates that there will be three post-positions and two three-letter words in a sentence of ten words.

Summarizing, the scholar Subramaniya Iyar made use of passive voice sentences in present tense. There will be six case markers, four nouns, three conjunctions and three postpositions and one pronoun in a sentence. The verb and word starting with vowel will be four-letter words with two or three syllables. The occurrences of more nouns reduce the occurrence of pronouns.

4. Conclusions

This study provides opportunities to introduce statistical techniques for identifying the special stylistic features and also for quantifying the writing styles of three Tamil scholars, namely, Mahakavi Bharathi, V. Kalyanasundram and Subramania Iyer using eighteen stylistic features. Articles written on India's Freedom Movement by these scholars are considered for this study. Bharathi had written sentences in past tense with the least function words. V. Kalyanasundram is identified as a writer who has used maximum number of function words in active voice sentences with past tense. The third writer, Subramaniya Iyar has written sentences in passive voice but in present tense and is not identified with any distinct stylistic features.

References

Bailey R.W. (1979). The Future of Computational Stylistic. *Association for Literary and Linguistic Computing Bulletin*.

Burrow J. (1987). Word-Patterns and Story-shapes: The Statistical Analysis of Narrative Style. *Literary and Linguistic Computing*, vol. (2/2): 61-70.

Fucks W. (1952). On the Mathematical Analysis of Style. *Biometrica*, vol. (39): 122-129.

Herdan G. (1941). *The Advanced Theory of Language as Choice and Chance*. The Hegue.

Herdan G. (1964). On Communication between Linguistics. *Linguistics*, vol. (9): 71-76.

Holmes D.I. (1985). The Analysis of Literary Style: A Review. *Journal of the Royal Statistical Society*, Series A (155): 91-120.

Holmes D.I. and Forsyth R.S. (1995). The Federalist Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing*, vol. (10): 11-27.

Holmes D.I. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, vol. (13): 111-117.

Johnson N.L. and Kotz. S. (1967). *Discrete Distributions*. Houghton Miffin Company Boston.

Lawley P.A. and Maxwell. A.E. (1971). *Factor Analysis as a Statistical Method* (2nd Ed.). American Elsevier Publishing and Co. New York.

Manimannan G. and Bagavandas M. (2001). The Authorship Attribution: the case of Bharathiyar. In *Paper Presented at National Conference on Mathematical and Applied Statistics*, Nagpur University.

Mealand D. (1997). Measuring Genre Differences in Mark with Correspondence Analysis. *Literary and Linguistic Computing*, vol. (12/4).

Miles J. and Selvin H.C. (1966). A Factor Analysis of the Vocabulary of Property in the Seventeenth Century. In Leed J. (Ed.), *The Computer and Literary Style*. State University Press.

Mosteller F. and Wallace D.L. (1964). *Applied Bayesian and Classical Inference, The Case of the Federalist Papers*. Addition-Wesley, Reeding.

Peng R. and Hengartner N. (2001). *Quantitative Analysis of Literary Style*. University of California, CA90095.

Peng R. and Hengartner N. (2001). *Statistical Aspects of Literary Style*. Yale University, CC-99.

Stylistic features	Abbre viations	Mean values			Two samples t-statistic values			Euclidean values		
		MB	VK	SI	MB-VK	MB-SI	VK-SI	MB-VK	MB-SI	VK-SI
Noun	P_Noun	34.26	45.47	41.8	09.38	06.21	02.91	125.49	56.73	13.47
Intensifier	P_Int	00.05	06.07	06.1	11.63	10.31	00.09*	31.07	31.79	0.00
Infinitive	P_Inf	00.58	01.33	03.6	02.69	06.14	04.26	0.57	9.18	5.19
Pronoun	P_Pro	07.72	07.73	06.6	00.01*	01.55*	01.64*	0.00	1.19	1.20
Tense	Tense	01.71	01.77	01.4	01.32*	05.84	08.42	0.00	0.08	0.11
Numeral	P_Nume	03.99	05.27	05.3	02.26	02.20	00.13*	1.62	1.83	0.01
Two-Letter Word	P_Two	10.61	09.79	09.5	01.04*	01.31*	00.31*	0.68	1.13	0.06
Three-Letter Word	P_Thre	19.24	20.18	18.3	00.86*	00.69*	01.52*	0.88	0.78	3.32
Four-Letter word	P_Four	20.46	25.66	25.9	04.20	03.86	00.19*	2.08	30.0	0.07
Word starting with vowels	P_Vowe	27.74	33.36	28.8	04.09	00.72*	03.35	31.61	1.27	20.22
Verb	P_Verb	23.43	21.40	23.6	02.52	00.22*	02.19	4.13	0.06	5.18
Voice	Voices	02.25	01.55	02.1	10.88	01.69*	08.69	0.49	0.01	0.34
Syllable	P_Sylla	151.10	119.70	163.0	02.82	00.88*	03.40	989.31	161.21	1949.20
Post position	P_Post	13.43	35.26	31.3	15.00	11.68	02.06	476.38	322.53	14.95
Clitics	P_Clitic	14.14	34.25	33.4	16.31	14.93	00.52*	404.28	371.69	0.68
Case marker	P_Case	38.65	69.95	63.0	15.15	12.45	02.75	980.00	596.68	47.30
Adverb	P_Adverb	04.39	02.26	02.8	04.22	02.78	01.29*	4.54	2.37	0.35
Conjunction	P_Conjun	22.65	42.15	35.5	11.11	07.42	03.48	380.24	165.74	43.90
Total								3458.37	1754.2	2105.60
SQRT								58.81	41.88	45.89

* not significance at 5% level

Table 1. Mean value, Two-samples t-statistic values and Euclidean distance values

Factors	MB			VK			SI		
	Eigen values	Percentage of variance	Cumulative percentage	Eigen values	Percentage of variance	Cumulative percentage	Eigen values	Percentage of variance	Cumulative percentage
1	2.223	12.3	12.3	2.018	11.2	11.2	2.492	13.8	13.8
2	1.560	8.7	21.0	1.645	9.1	20.3	1.590	8.8	22.7
3	1.420	7.9	28.9	1.463	8.1	28.5	1.483	8.2	30.9
4	1.270	7.1	36.0	1.329	7.4	35.9	1.217	6.8	37.7
5	1.167	6.5	42.4	1.150	6.4	42.3	1.135	6.3	44.0
6	1.103	6.1	48.6	1.068	5.9	48.2	1.094	6.1	50.1
7	1.035	5.7	54.3	1.048	5.8	54.0	1.008	5.6	55.7
8	1.013	5.6	59.9	1.004	5.6	59.6	0.966	5.4	61.0
9	0.961	5.3	65.3	0.943	5.2	64.8	0.903	5.0	66.0
10	0.945	5.3	70.5	0.916	5.1	69.9	0.876	4.9	70.9
11	0.856	4.8	75.3	0.843	4.7	74.6	0.829	4.6	75.5
12	0.837	4.6	79.9	0.784	4.4	78.9	0.772	4.3	79.8
13	0.758	4.2	84.1	0.745	4.1	83.1	0.720	4.0	83.8
14	0.702	3.9	88.0	0.738	4.1	87.2	0.673	3.7	87.6
15	0.646	3.6	91.6	0.685	3.8	91.0	0.658	3.7	91.2
16	0.546	3.0	94.7	0.580	3.2	94.2	0.597	3.3	94.5
17	0.450	2.8	97.4	0.540	3.0	97.2	0.538	3.0	97.5
18	0.460	2.6	100.0	0.502	2.8	100.0	0.447	2.5	100.0

Table 2. Factor analysis-Initial statistics

Factors	MB	VK	SI
FACTOR 1	(.75105) P-VOWE (.67532) P-VERB (.55204) P-TWO (.58042) P-THRE (.66074) P-FOUR	(.65782) P_CASE (.56787) P_POST	(.71539) P_VERB (.54890) P_FOUR (.64105) P_SYLLA (.34474) P_VOWE
FACTOR 2	(.83945) P_CLITIC (.85563) P_CASE	(-.53092) P_NOUN (.73761) P_NUME (-.68164) P_VERB	(.77953) P_CASE (.76640) P_CLITIC
FACTOR 3	(.70597) P_NOUN (.82167) P_PRO	(.75332) P_FOUR (-.67378) P_THRE (.59352) P_TWO (.75897) P_PRO	(-.66287) P_PRO (.62711) P_NOUN
FACTOR 4	(-.66491) TENSE (.69586) P_NUME	(.79795) P_ADVERB (-.39408) TENSES (.83215) P_SYLLA (.66924) P_CONJON (.71794) VOICE (.44038) P_VOWE (.73773) P_INF (.46984) P_INT (.56534) P_CLITIC	(-.71236) VOICE (.59945) P_CONJON
FACTOR 5	(.73510) VOICE (.62676) P_POST		(.79019) P_TWO (.56156) P_ADVERB
FACTOR 6	(.66129) P_SYLLA		(-.76712) TENSE
FACTOR 7	(.18315) P_INT (.66689) P_INF (.57019) P_ADVERB (-.52044) P_CONJON		(.10776) P_INF (.61875) P_NUME (.39798) P_POST (.76640) P_CLITIC (.72258) P_THRE

Factor scores are given in the brackets

Table 3. Grouping of stylistic features according to factor scores

A Text Mining Strategy based on Local Contexts of Words

Simona Balbi, Emilio Di Meglio

Dip. di Matematica e Statistica – Università "Federico II" di Napoli

sb@unina.it, edimegli@unina.it

Abstract

Aim of the paper is to propose a Text Mining strategy based on statistical tools, which make more efficient the extraction of information buried in massive quantities of documents. Usually, in Text Mining procedures (such as in textual data analyses) we deal with a *corpus* consisting of a set of documents. In order to build the data structure to be processed, each document is encoded in a *document vector*, according to the *bag-of-words model*, which associates words and their frequencies for the given document. Documents are considered as a whole. The proposed mining strategy identifies *interesting* sentences in the *corpus* we deal with, where to concentrate the knowledge extraction. The sentence interest will depend on the researcher's objective. The proposed procedure is useful when we are interested in local contexts for words. Prior information, i.e. expert knowledge, is included, as an input for the procedure, but differently to content analysis, the key-word system is automatically built. The strategy can be applied in any case we can introduce information for partitioning documents in lower order grammatical units (e.g. sentences, but also paragraphs, etc.). The mining procedure consists in two steps: first of all the Text Categorisation, i.e. the recognition of the *interesting* sentences, by means of a statistical segmentation procedure, and then the knowledge extraction from the identified sub-texts. The procedure first step produces association rules useful in filtering e-mail, chat, or Web access, too. The paper aims at contributing to the day-by-day wider literature on Text Mining, devoted to go beyond the "bag-of-words" model of structuring the data set in document vectors, enhancing the role of a statistical perspective. An application on Italian on-line job offers ends the paper, showing the effectiveness of the proposal.[1]

Keywords: bag-of-words, association rules, segmentation, text categorisation.

1. Introduction

It is well-known that while data mining deals with numerical data arranged in structured data bases, Text Mining deals with unstructured documents, written in natural language. There are not negligible consequences in dealing with unstructured materials. In numerical data bases, the elementary units are the field contents in records. In Text Mining, things are not so simple. Not trivially, from a statistical viewpoint, it is not definitely clear how to structure the data set to be analysed. First of all we need to answer the questions on which are the statistical units, and which are the variables.

In the field of linguistic statistics and textual data analysis, we can find interesting applications working on letters (vowels, consonants), words (e. g. graphical forms, lemmas, textual forms), groups of words (e.g. repeated segments, quasi-repeated segments), sentences (i.e. grammatical self-contained speech units with a capital letter at the beginning and a full stop at

[1] The paper has been financed by the Italian Ministero dell'Università e della Ricerca scientifica, in the project OUTCOMES (Occupation as a University Target and Careers of Outgoing graduates Maximising their use of Educational Skills), PRIN 2002

the end), and so on. However, the complexity of the choice has some remarkable implications. We can structure the knowledge extraction process moving from data of different levels and orders, so as to make more efficient and effective our results. The practice of representing documents as bag-of-words (i. e. encoding documents in vectors, associating words and their frequencies, and processing documents as a whole) can be overcome, aiming at enhancing the context in which a word has been used.

Here we propose a methodology, based on Text Categorisation, which exploits the structural organization of documents in sentences. This strategy consists of two step: in the first step, statistical units are sentences described by words, while in the second step units are words, characterised by their frequencies in each sentence. In the first step, Text Categorisation is undertaken to discriminate interesting sentences from uninteresting ones (having as an input expert knowledge). This approach, given an informative need, allows to eliminate all the information not useful to satisfy this specific need. Once we have reduced the *corpus* only to the sentences related to our problem, we can apply a proper statistical method for knowledge extraction. In this way, computational speed ups are obtained; documents and terms similarities (fundamental measures for any Text Mining application) are targeted on the particular informative need. This strategy attempts to go beyond the bag-of-words model. In fact, being the bag-of-word built on sentences, relations among words are, in some way, taken into account.

For showing the effectiveness of the proposal, two different typologies have been built, starting from the same collection of on line job offers: one on the original corpus and the other one on the reduced corpus obtained by discarding sentences not containing skill requirements. Comparing results gives an idea of the information gained, as consequence of this light contextualisation.

2. Text Categorisation

Text Categorisation (also named Text Classification) is a Text Mining task with a broad domain of applications, ranging from automatic document indexing to document filtering, metadata generation, word sense disambiguation, hierarchical organization of web documents and any application that require selective documents organization, such as limited Web access (e.g. children, etc.). Text Classification is therefore of great utility in business and Information and Communication Technology applications but is also useful to extract knowledge that constitutes the starting point for other Text Mining applications.

Text Classification can be defined as the task of assigning a Boolean value to each pair (d_i, c_i) of DxC where D is a set of documents and C is a set of pre-defined categories. A value True is assigned if document d_i is classified under category c_i and a value false is assigned if document d_i is not classified under category c_i . Formally, the task is to approximate an unknown target function $\varphi: DxC \rightarrow \{T,F\}$ that describes how documents should be classified with a function $\hat{\varphi} : DxC \rightarrow \{T,F\}$ called the *classifier* in such a way that $\hat{\varphi}$ "coincides as much as possible" with φ.

The degree of "coincidence" between the target function and the classifier determines the effectivity of the classification algorithm. The classifier is built on a training set and is validated on a test set. The general problem of statistical classification in textual domain can be summarized as follows. We have a training set of document vectors, each labeled with one class, called target value. In practice, each object of the training set is represented in the form (x_i , y_i) where x_i is the document vector and y_i the class label value. The goal is to learn a

mapping or a function f(x) that is able to predict a class value y, given a document (Hand *et al.*, 2001).

At this point it is useful to define the concepts of *model class*, *score function* and *optimization strategy*. The *model class* is a parametric family of classifiers, that is a function f(x, ξ), where ξ is a parameter vector. Usually, in Text Classification problems very few is known on this function. The *score function* is the function that numerically expresses the preference for a model over another. It is typically a function of the difference between the predicted value y^* and the true value y, measured with a proper dissimilarity index. The *optimization strategy* is the strategy used for finding the best parameters and models within the model class. The aim of the learning algorithm is therefore to minimize the score function as a function of ξ.

Once the parameters have been chosen, that is once the classifier has been trained, it is necessary to estimate its performance on a test set. The test set should never include data used in the training step. The model could in fact result overfitted. In literature, several techniques have been proposed. In the following section, the main techniques for classifying texts will shortly be reviewed.

Techniques for Text Classification

In general terms, we have three main families of statistical techniques for Text Categorisation: Decision Trees, regression methods and neural networks.

A decision tree is a tree structure that visualizes the document categorisation process. The tree is composed by nodes, branches and leaves. By following the tree from the top node it is possible to classify a new document by recursively choosing the appropriate branches until a leaf is reached. The tree is built on a training set with an iterative algorithm. The training set consists in a set of labeled documents, i.e. documents for which the category is *a priori* known. At each step it is chosen the variable which splits the data into groups leading to the greatest improvement of the score function. The iteration stops when each leaf contains one data point or identical data points or if a given stop rule is met. In this way the maximum tree is obtained. However, this tree describes perfectly the training set but leads to overfitting. For this reason the tree is pruned. Pruning consists in eliminating some branches of the tree in order to have a simpler model and a better performance. The goal is to find a model complex enough to capture the structures existing in the data, but not so complex to overfit. Most used algorithms are CART and C4.5.

Regression methods aim at explaining or predicting a continuous variable, on the basis of explicative variables. Linear Regression is the simplest and most widely used regression model. In Text Classification domain the aim is to obtain a binary value that indicates membership to a certain class. Logistic Regression, proposed for dealing with explicative categorical variables, allows to calculate a real valued ranking for the class membership. The result is, in fact, a *categorisation status value (CSV)*, i.e. a number between 0 and 1 representing the evidence for class membership. The parameters of logistic regression can be estimated using maximum likelihood. Regression methods give measures for the importance of each explicative variable in determining the class membership. It suffers however of some substantial drawbacks. First, it requires considerably more observations than variables to obtain valid estimates of parameters. It is also computationally expensive and this makes it a not suitable methodology for huge data sets.

A Neural Network text classifier is a network of interconnected computing units where the input units represent the terms; the output units, the categories and the weights on the

connections, the dependence relations among units. The classification process of a document is performed as follows: the term weights are loaded into the input units; the following units are activated on the basis of the input weights and the connections weights; unit activation is propagated forward in the network and the final classification decision result is determined by the value assumed by the output unit. Neural Network connections weights are learnt on a training set usually through a process called *backpropagation*. In backpropagation, the weights of a training document are loaded and if misclassification occurs the error is back-propagated so to adapt the parameters of the network to minimize this error. In practice, training a neural network consists in minimizing a score function in the parameter space, solving a non linear optimization problem. This is done by descending to a local minimum given a random starting point. The most used optimization techniques are *steepest descent* and *conjugate gradient*. The simplest type of neural network is the perceptron, which is a linear classifier. Non linear Neural Networks are instead networks with one or more hidden layers of units that represent, in Text Classification higher order relations between terms. This is one of the main advantages of using Neural Networks for Text Classification. Neural Networks are however computationally expensive if the network structure is too complex and cannot be as easily interpreted as classification trees and logistic regression.

This review does not claim to be exhaustive. Several methodologies have been proposed either in statistical literature either in Machine Learning literature, being Text Classification a problem studied in both fields. Among these we mention Support Vector Machines, linear discriminant analysis, Bayesian inference methods, maximum entropy modeling, genetic algorithms, association rules induction, bagging and boosting. For a complete overview see Sebastiani (2002).

3. A two-step strategy

In this paper we propose a two-step data analysis strategy, aiming at extracting knowledge from huge *corpora*, by both reducing the computational burden of the textual data analysis and accounting for the context in which words appear. In doing that we deal with some tasks and some solutions developed in Text Mining frame, first of all *Text Categorisation*.

As previously said, Text Categorisation consists in classifying text into one of several predefined categories. This task is performed by software called *automatic categorisers*. The software basically mimics the human process of evaluating the relevance of a document with respect to the topic of interest. A common way for doing that consists in introducing a keyword index, which defines logical rules, e.g. if a certain word occurs in a text, then the text will be identified in some sense and assigned to the related category. Words can be combined by logical operators (AND, OR, NOT), enriching the discriminative power of the rule. The procedure requires expert knowledge, i.e. the key words index, as input, or, alternatively, can be based on statistical methods for categorisation, as previously mentioned. Here we follow the latter approach, proposing statistical tools for reducing the quantity of external knowledge input and determining automatically the rules to apply in order to tag the texts. In the first step of the strategy, we apply the categorisation algorithm to discriminate part of documents (e.g. sentences), inherent to the objectives of the analysis. Therefore, only parts tagged as interesting will be consider in the following analysis step. The strategy can be useful in any case we can introduce information for partitioning documents in units of different levels. The synctatical structure presents in a document suggest to consider it as a complex unit, consisting of lower order units (sentences) and, hierarchically, much lower units (words). Moving at different levels will enable us to efficiently discover the knowledge we are looking for.

STEP 1

Aim: identifying sentences in the document, related to the topic of interest
Tools: statistical techniques for discrimination
Input: a training set and a test set, both consisting in sentences tagged by expert knowledge (0 = uninteresting; 1 = interesting)
Output: logical rules for identifying interesting sentences in the document to be analysed

STEP 2

Aim: eliminating uninteresting sentences in the document, by applying the logical rules identified in STEP 1;
Tools: advanced software or language for dealing with text (e.g. UltraEdit, Perl)
Input: the logical rules identified in STEP 1 and the document to be analysed
Output: a new document consisting of the sentences related to the topic of interest, to be analysed with the proper textual data analysis techniques, according to researcher's objectives

4. Looking for skills required in on-line Italian job offers

4.1. *Motivations*

A research net connecting eight Italian Universities is working on the correspondence between demand and offer of graduates on Italian labour market (*OUTCOMES: Occupation as a University Target and Careers of Outgoing graduates Maximising their use of Educational Skills*). Therefore it is important to analyse the skills required, and how they are required by the labour market (Balbi and Di Meglio, 2003). Here we propose to analyse on-line job offers, applying the proposed strategy. The first motivation is connected with some peculiarities of these texts. A job advertisement is usually composed by several sentences. Each sentence has a specific role, common in the most cases of job advertisements. Some sentences contain information on the required skills, others describe the professional profiles, others contain different information. In other terms, job advertisements have a light structure, and it is possible to identify which sentences are devoted to describe the required skills. Moreover, the vocabulary is not very wide, but some word connotations depend on the kind of sentences, where they are used. Therefore, a first analysis devoted to the mining of sentences has been performed.

4.2. *The first step: building logical rules for skill requirement sentences*

The document to by analysed is a set of job offers appeared on a specialised Italian site during the first 8 months of 2003. We have collected 2017 advertisements. We randomly choose a subset, in order to build the training set, consisting in 200 advertisements, composed by 670 sentences. After a manual tagging of the advertisements, we labelled 502 sentences with 0 (= *uninteresting*) and the remaining 168 with 1 (= *skill requirements*). Once the training set had been prepared, we went on building the logical rules for identifying skill requirement sentences in the rest of the document. For sake of simplicity, here we have chosen to ignore all the problems related to linguistics (e.g. lemmatisation, disambiguation, etc.), and we dealt with graphical forms (Lebart *et al.*, 1998). There are many statistical techniques devoted to classifying units (e.g. discriminant analysis, segmentation, and so on), and there are many mining algorithms developed for solving classification problems, by association rules. As suggested in a previous paper (Balbi and Summa, 2001), Symbolic Marking seems to efficiently perform in the case of Text Mining, being a segmentation method with a very high

performance in the case of huge data sets. Additionally its results are easily to be expressed in terms of logical operators.

4.2.1. *Symbolic Marking*

Symbolic Marking (SM, Gettler-Summa, 1998) is a non-binary segmentation technique which aims at finding the association structures in a group G_i belonging to a typology naturally defined, or obtained by a previous classification analysis. Symbolic Marking takes into account logical relations, as conjunctions and disjunctions, between attributes describing the units in G_i. The result can be expressed in natural language as logical rules, connecting attributes with logical operators.

The procedure finds *marking cores* g_k's, i.e. nuclei of individuals identical with respect to some variables, called characterising variables, y_j. The y_j's give a symbolic description of $g_k \in G_i$:

$$g_k : [y_1 = a_1] \wedge [y_2 = a_2] \wedge ... \wedge [y_r = a_r]$$

By joining n marking cores g_k's (defined by logical AND's, $\wedge$) with the logical disjunctive operator OR ($\vee$), one obtains a (partial) description of the group G_i, in terms of the selected characteristic variables y_j's, optimal according to the principles stated by Gordon (1999): *i)* minimising the number of false negatives (individuals in G_i, but not corresponding to the description); *ii)* minimising the number of false positives (individuals described by the conjunction of the marking cores, but out of G_i); *iii)* each conjunction of categories has to be statistically meaningful with respect to the test chosen for evaluating the strength of the link between G_i and each core (different measures have been proposed and used in the commercial software for symbolic marking, e.g. in SPAD a *test-value* based on a hyper-geometric distribution).

Thus the description is given by:

$$G_i : g_1 \vee g_2 \vee ... g_n$$

Two measures are associated to each marking core (together with the *test value*):

DEBOR = % of g_k's individuals belonging to g_k, but not to G_i

REC = % of G_i individuals described by g_k

The value of REC(g_k) can be cumulated (RECCUM) over the different cores and used to decide the number of cores to be considered.

Dealing with documents, we have a set of tagged documents described by its graphical forms, which play the role of the characterising variables, y_j.

4.2.2. *The extracted rules*

Applying the SPAD procedure for SM on the training set the following logical rules have been extracted.
Sentences containing skill requirements can be detected, by the rule:

conoscenza
OR (**esperienza *AND NOT* inquadramento**)
OR (**anni *AND NOT* responsabile *AND NOT* risorse**)
OR (**capacità *AND NOT* inquadramento**)

Sentences not containing skill requirements can be detected, by the rule:

***NOT* conoscenza *AND* (*NOT* anni) *AND* (*NOT* capacità)**

4.3. *The second step: discarding uninteresting sentences*

Once the logical rules have been identified, by means of macros built in UltraEdit software environment, only the sentences containing skill requirements have been selected.

A big reduction of the corpus size has been obtained, allowing a much faster computation in the subsequent analysis. From the original corpus containing 305,853 occurrences it has been extracted a sub-corpus (made of the original documents swept from the excluded sentences) containing 83,062 occurrences; circa 27% of the original corpus has therefore been selected.

4.4. *Knowledge extraction: a comparison of typologies*

The final aim of this analysis was to identify skill requirements in job offers. Once the text has been cleaned from all the redundant information, Text Mining techniques give results specifically tuned on the analysis scope. Here, a clustering procedure has been used in order to identify typologies of skills requirements. Typologies have been compared with the ones obtained on the original corpus.

Two lexical tables have been built: one on the original collection and the other one on the reduced collection. Following a common strategy in textual data analysis, preliminarily a Correspondence analysis has been performed on the two tables, and coordinates of documents on the first 10 factorial axes have been the input of the following clustering procedure.

By comparing results, the effectiveness of our proposal appears.

In the following table 1 the 5 classes identified in the full corpus are described. The optimal number of clusters is automatically identified by an index based on information loss.

Class 1 (74.9%)	**Class 2 (8.5%)**	**Class 3 (0.5%)**	**Class 4 (4.4%)**	**Class 5 (11.7%)**
Esperienza	Sede	Consulting	Master	Posti
Anni	Milano	Milano	Corso	Durata
Dati	Leader	Sede	Stage	Svolgimento
Curriculum	Cliente			
Capacità	Azienda			

Table 1. Typology on the full corpus

The clusters identified on the full collection are mainly described by information not useful for the analysis aim. Only the first cluster is partially described by words related to skills (*esperienza, capacità*). Classes 2 and 3 contain company descriptions and classes 4 and 5 descriptions of masters, internships and courses without a precise separation among these different offers.

Class 1 (85.5%	Class 2 (0.2%)	Class 3 (12.0%)	Class 4 (2.3%)
Esperienza	Esperto	Word	Dinamici
Anni	Relazioni	Excel	Vendita
Titolo	Industriali	Internet	Agenti
Disponibilità	Milano	Access	Obiettivi
Doti		Stage	Chimica
		Laurea	

Table 2. Typology on reduced corpus

The typology of reduced collection (Table 2) describes four skill groups: the first is a group mainly characterized by *esperienza*. This class describes the skills required to experienced workers; these skills are very heterogeneous even if they are all characterized by experience in some working field further described with other words in the job offer. Being this group also quite large a deeper investigation is needed. The other three describe more defined skill profiles, respectively: industrial relations experts, internship candidates and salesmen. In particular we see that internship candidates are required to have a university degree and to have basic computer knowledge. To salespersons is instead required to be dynamic and to work by objectives. This typology, differently from the first one reaches the analysis objectives describing the skills required. An overview of the skills can be obtained by a finer clustering.

In the 9 class typology of Table 3 we get a complete overview of the different skills required by the job market. Classes 2,3,4 of the previous typology correspond to classes 7,8 and 9 of the new one, while class 1 is split in 6 new classes. The first describes a general profile for which knowledge, computer, foreign languages, dynamism and interpersonal skills are important. Class 2 describes jobs in the field of distribution for which are important degree, residence and interpersonal skills. Class 3 describes the graduates in economics for which age, experience and leadership are important. In Class 4 we find engineers profiles to which projecting and drawing skills are required. In class 5 there are profiles of professional salesman with motivation and experience. Class 6 is a residual group of advertisements offering computer courses to workers.

5. Conclusions

Here we have proposed a methodology, based on the principles of Text Classification, meant to extract interesting (in a specific domain) patterns of words. In this proposal there is a first attempt to go beyond the traditional bag-of-words as we encode sentences and not the whole documents. In this way, some sequential boundaries are built, as we are not interested in the general contexts in which words are used but in the local contexts, conveying the specific information of interest. This is a consequence of the fact that most of the times, documents are not completely unstructured; let's think about scientific articles (abstract, introduction, proposal, conclusion). This structural division can be more or less evident. Aim of this strategy is to exploit the structure embedded in the document and improve Text Mining task performances. Job offers are a very well suited example of semi-structured documents and we used an Italian on line job offers collection in order to show the effectiveness of our proposal, in clustering skill requirements.

Job advertisements are usually composed by the description of the job profile, the required skills, and further information (e.g. selection process, retribution, contacts). In the first step of our procedure we dealt with a tagged training set, which differently labeled skill requirement sentences. This first step can be viewed as a pre-processing step, aiming at identifying the words, and the association among words, which discriminate sentences carrying information about the selected topic. Its output are rules, validated on a test set. Then, rules are applied to the entire corpus, in order to discard not interesting parts. A reduced corpus is produced, input for further mining processes. The extracted subtexts in fact have less variability and contain less noise. This leads to better performances of visualization, retrieval and clustering techniques, as shown by the job offers.

Class 1 (25%)	**Class 2 (1,9%)**	**Class 3 (35,1%)**	**Class 4 (10,6%)**	**Class 5 (12,2%)**
Conoscenza	Diploma	Laureato	Meccanica	Professionisti
Office	Residenza	Economia	Disegno	Vendita
Lingue	Distribuzione	Esperienza	Tecnico	Laureati
Windows	Interpersonali	Giovane	Ingegnere	Motivazione
Interpersonali		Leadership	Automotive	Esperienza
Dinamismo		Responsabilità	Progettazione	Agenti

Class 6 (0,9%)	**Class 7 (0,2%)**	**Class 8 (11,9%)**	**Class 9 (2,2%)**
Corso	Esperto	Word	Dinamici
Windows	Relazioni	Excel	Vendita
Diplomati	Industriali	Internet	Agenti
Laureati	Milano	Stage	Obiettivi
		Access	Chimica

Table 3. Typology on reduced corpus with nine clusters

References

Balbi S. and Di Meglio E. (2003). Text Mining on-line job offers. *Bulletin of the International Statistical Institute, 54th session,* vol. (60/1): 65-67.

Balbi S. and Gettler-Summa M. (2001). Identifying Lexical Profiles by Symbolic Marking. In *Book of Short Papers CLADAG2001*, Palermo: 185-188.

Gettler-Summa M. (1998). *MGS in SODAS: Marking and Generalization by Symbolic Objects in the Symbolic Official Data Analysis Software*. Cahier 9935. Université Dauphine LISE CEREMADE.

Gordon A.D. (1999). *Classification*. Chapman & Hall CRC.

Hand D., Mannila H. and Smyth P. (2001). *Principles of Data Mining.* MIT Press.

Lebart L., Salem A. and Berry L. (1998). *Exploring Textual Data*. Kluwer Academic Publishers.

Manning C.D. and Schütze H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Sebastiani F. (2002) Machine Learning in automated Text Categorization. *ACM Computing surveys,* vol. (34/1): 1-47.

Simple linguistic methods for improving a word alignment algorithm

Ana-Maria Barbu

Research Institute for Artificial Intelligence (RACAI) – Calea 13 Septembrie 13, PC 050711, Bucharest – Romania
abarbu@racai.ro

Abstract

This paper approaches word alignment problems and aims to show how some linguistic methods can combine with the statistical ones in order to get better results. In this sense, the system that participated in the shared task HLT/NAACL 2003 is presented in its actual version including the improvements the paper focuses on. This system, called TREQ-AL, works on parallel bilingual texts and needs some common pre-processing, such as tokenization, sentence-alignment, tagging, which are also shortly presented. In order to fulfill the word-alignment task, TREQ-AL relies on the output of a statistical translation-equivalence extractor along with which is described in a main section of the paper. The linguistic improvements taken into account are shown in another section and they refer to the cognate detection, the precedence constraint, pair assignments, collocations and language-specific rules. Instead of conclusions, graphical evaluation data are offered in the final section of the paper.

Keywords: word alignment, translation equivalence, linguistics, statistics.

1. Introduction

The work this paper relies on was roughly developed in the shared task organized as part of HLT/NAACL 2003 workshop on "Building and Using Parallel Texts: Data Driven Machine Translation and Beyond". The task consisted of finding correspondences between words and phrases in the parallel texts of an aligned bilingual corpus. Assuming sentences in language L1 translated into language L2, the task supposed to build a word alignment system indicating which word token in the text of language L1 corresponded to which word token in the text of language L2. Corpora for two pairs of languages were provided: French-English (20 million words) and Romanian-English (1 million words) in order to train and test the competitive systems. All data were pre-tokenized. Final results should indicate the number of each parallel sentence and pairs of numbers corresponding to numbered tokens (both words and punctuation marks) associated in the two texts of that sentence. Tokens not translated were associated with nulls. More details about this evaluation exercise can be found in Mihalcea and Pedersen (2003).

Our team participated in Romanian-English subtask, beside other six teams from around the world, with the system called TREQ-AL (for TRanslation-EQuivalence ALigner). The complex work devoted to this task is partially described in Tufis *et al.* (2003). This system has been built in less than three weeks and in despite of this short time the results were quite comparable with those of the most powerful systems. Since that moment we have improved the word-alignment algorithm especially by developing its linguistic component. Our new results, better than those of the other participant systems in the workshop—either purely linguistics-based system like Zhao and Vogel (2003) or purely statistics-based ones like Dejean

et al. (2003), have confirmed our intuition that linguistic methods can complete the statistical ones.

Figures can clearly express the improvement we talk about, if we take a look at the values of three main measures: Precision (P)(which indicates how many alignments are correct from the total alignments offered by the system), Recall (R) (which indicates how many correct alignments are found from the total alignments that should be found) and F-measure (F) (which is defined as 2*P*R/(P+R) and is a "correlation" of the first two). For the version of TREQ-AL at the moment of the shared task, we have got the following values: P=81.29%, R=60.26 and F=69.21, for non-null alignments. After developing the algorithm the measures amounted to significantly better values: P=84.95%, R=65.13 and F=73.73.

The paper aims at describing the ways the mentioned growth was possible and, in general, how linguistics can help statistics in natural language processing, in the particular application of word-alignment for bilingual texts. The content of the paper is structured on the following sections. The first one shortly presents the basic steps for processing parallel texts. A main section is devoted to the statistical part of the task, as a primary treatment; more precisely, it describes the central algorithm TREQ for translation equivalents extraction, which is completed, afterwards, with non-linguistic techniques of word to word alignments. The second main section emphasizes linguistic corrections made to improve the results. This kind of adjustments relies on the syntactic similarities displayed by Romanian and English. Some of these are very general, so that they can be inferred, at least, for some other Romance and Germanic languages. Paper ends with conclusions about gains of the approach.

2. Pre-processing steps of a parallel text

A parallel text is an association of two (or more) texts in different languages, which represent translations of each other. In order to get the most informative parallel text, appropriate for complex and efficient treatments, raw texts need to pass through more steps, described below.

2.1. *Segmentation*

A rough computational definition for 'word' is, as regarding written texts, any item separated by white spaces. However this concept is not operable in natural language, in that there are items, for instance in English, like "he's" (for *he is*) or "couldn't" (for *could not*) which do not represent only one word, but two. Without doubt, there are such elisions in many languages and they have to be splitted (by white space insertions) for a good machine understanding. Moreover, punctuation is conventionally written next to the previous item, without any space. What results does not form one word, as it can not be found in any lexicon. So punctuation marks have to be splitted, too.

On the other hand, there are a lot of multiword lexical tokens in any language. English verbs with particle (e.g. *cut off*, *set on*) or grammatical phrases (e.g. *in despite of*) are good examples in this sense. All these words with one lexical meaning have to be treated as a compound unit and marked as such. Usually, this is done by replacing spaces with underscores (e.g. *in_despite_of*). This task is called segmentation (or tokenization) and the program that performs it, segmenter (or tokenizer).

Note that a good delimitation of lexical tokens is important for translation purposes and, in our case, for word alignment, where 'word' has just the meaning of lexical token. For our work, we use Philippe di Cristo's multilingual segmenter Mtseg (http://www.lpl.univ-aix.fr/projects/multext/MtSeg) developed for MULTEXT project.

2.2. *Sentence alignment*

After tokenizing the two parts of a parallel text, each sentence of them gets a unique identifier. Then texts and sentence identifiers are given as input to the sentence aligner, whose task is to put in correspondence one or more sentences in a language with one or more sentences in the other language, through their identifiers. Each correspondence marks an alignment unit, also called from now on *translation unit*. Such a sentence aligner is CharAlign, fully described in Gale and Church (1993).

Obviously, the sentence alignment step is crucial for word-to-word alignment, due to the fact that the latter is only a refinement of the former. Errors in sentence alignment can trigger even more errors at word alignment level. That is why the results of this process should be checked out. A verification criterion could be the percentage of alignment units consisting of correspondence of one sentence to only another one, which should be more than 94%, as we have noticed in our experiments on a parallel corpus of six languages. So just those translation units that do not display one-to-one correspondence could hide alignment errors.

2.3. *Tagging and Lemmatization*

Another very useful step for natural language processing is that of indicating the proper part of speech for each lexical token. Sometimes, more parts of speech are appropriate for the same word, in the case of homographs. In English, there are plenty of homographs, so for example *well* appears in lexicon as an adverb, adjective, noun and verb. In text, *well* has to get only one part of speech, depending on its context. This is the task of a tagger. So tagging is the process of assigning a morpho-lexical label to each token and of solving ambiguity cases.

First of all, taggers need a training stage in which they learn, on the one hand, words and their appropriate morpho-syntactic tag(s) and, on the other hand, sequences of such tags as they are encountered into pre-tagged texts. With these data, taggers build a *language model* and afterwards they apply it to unknown texts.

Tagger performances depend on the size and the quality of the tagset, in that the morpho-syntactic labels have to be chosen so that they are not too many and offer relevant grammatical information. For our work, we use TnT (Brants, 2000), a trigram HMM tagger, whose accuracy amounts to more than 99%, for a Romanian tagset of 92 non-punctuation tags.

Lemmatization is a process of finding the normal form (lemma) for each word, especially for the inflected ones. That can be done either by program or, the simplest, by looking up into a monolingual lexicon of inflected forms. Working on lemmas is very important in natural language processing from statistical point of view: the number of wordforms submitted to statistical methods decreases and that of their occurrences grows.

In the example below, one can see the result of the pre-processing steps described until now for an one-sentence parallel text in *xml* format. Note that the element *tu* marks a translation unit, *seg* -the language, *s* -a sentence, *w*- a wordform, *c* -a punctuation mark, while the attribute *id* specifies the sentence/translation unit identifier, *lemma*—the normal form, and *ana*—the morpho-syntactic label.

Exp. 1
```
<tu id="Ozz.60">
<seg lang="ro"><s id="Oro.60"><w lemma="de_exemplu" ana="R">de_exemplu
</w> <c>,</c> <w lemma="ca" ana="C">ca</w> <w lemma="istoric" ana="N">
istoricul</w> <w lemma="Treptow" ana="N">Treptow</w> <w lemma="fi" ana="V">
este</w> <w lemma="american" ana="A">american</w> <c>.</c> </s></seg>
<seg lang="en"><s id="Oen.60"><w lemma="that" ana="Di">that</w> <w
lemma="historian" ana="N">historian</w> <w lemma="Treptow" ana="N">Treptow
```

```
</w> <w lemma="be" ana="V">is</w> <w lemma="an" ana="Di">an</w> <w
lemma="American" ana="A">American</w> <c>,</c> <w lemma="for_example"
ana="R">for_example</w> <c>.</c> </s></seg>
</tu>
```

3. From TREQ to TREQ-AL

3.1. *TREQ system*

The TREQ system requires sentence-aligned parallel text, tokenized, tagged and lemmatized, as we have already discussed. Its aim is to get translation equivalences from the text, in other words to create a bilingual dictionary based on that text.

The algorithm makes use of two underlying assumptions:

1. a lexical token is translated by only one token in the other language (Melamed, 2000);

2. the two members of a translation equivalence have the same part-of-speech.

These assumptions are restrictive, indeed, but they do not prevent additional processing units from recovering some of the missed or incomplete translations, as we shall see in TREQ-AL's case.

The baseline of TREQ is quite simple (for a complete presentation see (Tufis & Barbu, 2002)). For each translation unit, every word (actually, its lemma) is paired with every word of the same part-of-speech in the other language. Thus, lists of translation candidates result for each part-of-speech. For instance, the lists extracted from the example 1, in the previous section, are given below:

Exp. 2
```
N (noun): (istoric historian), (Treptow historian),
                (istoric Treptow), (Treptow Treptow)
  V (verb): (fi be)
  A (adjective): (american American)
  R (adverb): (de_exemplu for_example)
```

Note that the words without correspondent in the other language are ignored (e.g. Di: *that* and *an* –in English, and C: *ca* –in Romanian).

Let a pair be noted $p=(w_1\ w_2)$. Then the algorithm computes, as if there are more parallel texts only with nouns, verbs etc., the score for each pair p with the following loglikelihood formula:

$$LL(w_1\ w_2) = 2 * \sum_{j=1}^{2} \sum_{i=1}^{2} n_{ij} * \log \frac{n_{ij} * n_{**}}{n_{i*} * n_{j*}}$$

where,

n_{11} = the number of p's occurrences in the whole text;
n_{12} = the number of pairs in which there is w_1 but not w_2, i.e. $(w_1\ \neg w_2)$;
n_{21} = the number of pairs in which there is w_2 but not w_1, i.e. $(\neg w_1\ w_2)$;
n_{22} = the number of pairs in which there is neither w_1 nor w_2, i.e. $(\neg w_1\ \neg w_2)$;
n_{1*} = the number of pairs in which there is w_1 (irrespective of its associations);
n_{*1} = the number of pairs in which there is w_2 (irrespective of its associations);
n_{2*} = the number of pairs in which there is w_1 does not appear;
n_{*2} = the number of pairs in which there is w_2 does not appear;
n_{**} = the total number of pairs.

After the score-calculating stage, the translation units are inspected again, one by one, and for each of them the candidates with the highest scores are chosen as translation equivalences if the scores are higher than a confidence threshold (empirically set to 9). Note that if the pair p

wins, other pairs containing w_1 or w_2 are ignored. Thus, the TREQ dictionary is made up of all the selected translation equivalences taken once.

3.2. *TREQ-AL system*

TREQ-AL has as input the TREQ lexicon and the parallel text to be aligned at the word level. The alignment is expressed, this time, in word-position terms, that is, the words are represented by their position in the translation unit, separately for each language. For the example 1, retaken here simplified and with words numbered, TREQ-AL should produce the indicated list of assignments (where '–1' means that the corresponding word is not translated):

Exp. 3 **RO:** 0>de_exemplu 1>, 2>ca 3>istoric 4>treptow 5>fi 6>american 7>.
EN: 0>that 1>historian 2>treptow 3>be 4>an 5>American 6>, 7>for_example 8>.
word alignment: (0 7), (1 6), (2 0), (3 1), (4 2), (5 3), (6 5), (7 8), (-1 4)

In order to get such results, TREQ-AL goes through the following processing steps reiterated with each translation unit.

3.2.1. *Dictionary looking-up*

First, for each word in the source language (here Romanian), TREQ-AL looks for the appropriate translation equivalent(s) into the TREQ lexicon. For those words that are not found in the lexicon, the system searches cognates (that is, similar words with the same meaning in the two languages) among not assigned target words. The looking-up is done irrespective of the part-of-speech, in order to avoid tagging errors. For instance, again in the example 1 (section 2.3), the English word *that* (position 0) is wrongly tagged as determiner (Di), and usually it could not be aligned with the Romanian conjunction *ca* (position 2). However, TREQ produces the conjunction pair (*ca that*) from many other occurrences in the parallel text and therefore this assignment can be recovered.

This step results in a list of (possibly non-consecutive) positions of those source words for which one or more translation equivalents were found, as the example 4 shows in the first column.

Exp. 4

RO: 0>sa 1>sine 2>aplica 3>lege 4>si 5>in 6>caz 7>un 8>apropiat 9>al 10>domn 11>talpes 12>, 13>si 14>al 15>nu 16>mai 17>sti 18>cine 19>!

EN: 0>that 1>the 2>law 3>be 4>also 5>enforce 6>in 7>the 8>case 9>of 10>a 11>person 12>close 13>to 14>mr 15>talpes 16>and 17>to 18>who 19>know 20>who 21>else 22>.

2- aplica /enforce-5*	2-5 aplica /enforce-5	2-5 aplica /enforce-5
3- lege /law-2/be-3/enforce-5	3-5 lege /law-2/be-3/enforce-5	3-3 lege /be-3
4- si /that-0/and-16	4-0 si /that-0/and-16	4- -1 si /*
5- in /in-6/of-9/to-13/to-17	5-6 in /in-6/of-9/to-13/to-17	5-6 in /in-6
6- caz /case-8	6-8 caz /case-8	6-8 caz /case-8
7- un /that-0/a-10	7-10 un /that-0/a-10	7-10 un /a-10
8- apropiat /close-12	8-12 apropiat /close-12	8-12 apropiat /close-12
11- talpes /talpes-15	11-15 talpes /talpes-15	11-15 talpes /talpes-15
13- si /that-0/and-16	13-16 si /that-0/and-16	13- -1 si /*
16- mai /also-4/else-21	16-21 mai /also-4/else-21	16-21 mai /else-21
17- sti /be-3/know-19	17-19 sti /be-3/know-19	17-19 sti /know-19
18- cine /that-0/who-18/who-20	18-20 cine /that-0/who-18/who-20	18-20 cine /who-20
Dictionary looking-up	**Up-bottom alignment**	**Bottom-up alignment**

*The alignment-line structure: ROposition – ENposition ROword /ENword1-ENposition1/…

3.2.2. *Up-bottom alignment*

The next step after dictionary looking-up is the up-bottom (or left-to-right) alignment, which processes the text in its normal reading sense. The target of this step is to do primary assignments and to coarsely solve the translation ambiguity.

The way of choosing a target word w_j from an ambiguity list depends on three factors: the cognate status -*cog*, the positional distance to the previous assignment -d_a and the relative distance to the source position -d_r. So, a target position j wins if it gets the best (in general, minimal) value for one of these dimensions:

$$j \Leftrightarrow \min \{cog, d_a, d_r\}$$

In the cases of non-ambiguity, the unique translation equivalent represents the proper link.

Note that at the end of this step, a target position can be assigned to more than one source position if it satisfies the selection criteria, as one can see in the example 4 (EN-position 5 assigned to RO-2 and 3).

Another important task fulfilled in this step is the detecting of a*lignment chains*, that is, sequences of at least four consecutive words in the source part associated with consecutive or close to each other words in the target part. For instance, in the example 4 (the second column), the links (5 6), (6 8), (7.10), (8 12) are memorized as an alignment chain, because Romanian positions are consecutive and no distance between two successive EN-positions is bigger than 2. The alignment chains are of great confidence in the ulterior word alignment process.

3.2.3. *Bottom-up alignment*

This step tries to refine and correct the primary assignation. It achieves a bottom-up (or right-to-left) alignment and takes into account more information than the previous step. The alignment criterion is a function depending on the following data:

- the distance to the lower assignment;
- the distances to the upper two assignments;
- the distance between source positions (especially relevant in the cases of gaps);
- the a*lignment chains*;
- the precedence constraint (presented later).

The result is a strict one-to-one word mapping, which can reflect modifications or even deletions (marked in the example 4 with '-1') of the links in the previous step, if no translation equivalent satisfies the alignment criterion. Note that this criterion affects both the ambiguous and non-ambiguous positions.

The next two steps use general linguistic knowledge for aligning the words that remain unaligned because there is no translation equivalent for them or the existing one(s) missed the alignment criterion.

3.2.4. *Alignment zones*

The system delimitates and count off, in each part of the translation unit, contiguous pieces of text that begin with a conjunction, a preposition or a punctuation mark and end with the token preceding the next conjunction, preposition, punctuation or end of the sentence. These are used as alignment zones in that they are mapped from one language to the other via the links assigned in the previous steps. That helps to filter out the links that exhibit aberrant zone

mapping, for instance if the source words in zone 1 are aligned with target words in zones 2, 7 and again 2, then the link inducing the mapping with zone 7 is deleted. It should be said that it is possible to get some unmapped zones, namely those which contain no aligned words.

3.2.5. *The final word-alignment*

Now, the algorithm looks for aligning un-linked words inside the zones mapped at the previous step. First, the words of the same part-of-speech are aligned and then the system tries to do cross-part-of-speech or multiple alignments according to some language-specific rules.

For an unmapped zone, the search space for new alignments is that between the closest links on both sides of that zone.

Any word in a language or another that has not been aligned after these processing steps is automatically assigned a null link.

4. Linguistic improvements for TREQ-AL

In order to fill the gaps in the word alignment and to solve ambiguity classes, beside the procedure presented until now, we applied the following linguistic assumptions and methods.

4.1. *Cognate detection*

Even if TREQ has a special module for treating cognates, for statistical reasons not all the cognates in the parallel text are validated as translation equivalences. Such a reason could be a very low number of occurrences in the text, which triggers a low statistical score not passing the necessary threshold. For recuperating this loss, TREQ-AL recalculates the cognate score for every possible translation pair not found in the dictionary, during the looking-up step. However some parts-of-speech, namely the functional ones, such as prepositions, conjunctions, pronouns, are ignored. The cognate-detection demarche turns out to be very benefic for the result of the word alignment, because an important number of translation equivalences are recovered. The linguistic base in the cognate detection is both the fact that many languages have plenty of words with common etymons, and the fact that, nowadays, there is an important terminology migration between languages.

A problem we had to solve was to set up the minimal limit for declaring words to be cognates. That has to be done so that the alignments gaps left by TREQ be filled with as many as possible correct equivalences. A too high threshold leaves many gaps unsolved, while a too low one induces alignment errors. From our experiments, we got the optimal cognate limit of 0.65. Without doubt, this limit depends on the language pair and the text nature. For illustration, we give below some Romanian-English pairs of cognates and their corresponding scores.

Exp.5

RO-word	EN-word	Cognate score
statistica	statistic	0.95
rominia	romanian	0.80
tipar	type	0.75
noiembrie	november	0.68
coruptie	corruption	0.65

4.2. *Precedence constraint*

Choosing the appropriate translation of a word from a list of many possible translations is an important step, because an error at this level triggers errors in other points of the sentence. In order to help the disambiguation process we set on the precedence constraint, relying on the general linguistic fact that certain parts-of-speech always precede others. For instance, prepositions, articles, determiners, even conjunctions precede nouns, adjectives, adverbs and sometimes verbs (especially participles).

At the first reading of the bilingual text, sequences of positions in the sentence are memorized if they start with prepositions, articles, determiners or conjunctions, and contain nouns, adjectives, adverbs and verbs only. That simulates somehow a chunker task for prepositional and noun phrases and exploits the fact that conjunctions (either coordinating or subordinating ones) always precede conjuncts.

The precedence constraint applies at the level of bottom-up inspection. Its role is of preventing the choosing of a translation from an ambiguity list if the order indicated by the memorized sequence is not respected, or of giving priority to the positions indicated in the sequence. In the apparently trivial example below, given the (preposition-noun) sequences 3-4 in Romanian and 5-6 in English, and the assignation of the Romanian position 4 to the English position 6 (done by TREQ), there is no difficulty in choosing, with priority, the English position 5 from the ambiguity `/on-3/on-5/in-13` as assignment for the Romanian position 3.

Exp. 6 **RO:** 0>poate 1>ca 2>intimplare 3>de 4>vineri 5>seara 6>de_la 7>otv 8>el9>vrea 10>determina 11>pe 12>cel 13>mai 14>mult 15>sa 16>exclama 17>exaspera 18>: 19>ajunge 20>!

EN: 0>maybe 1>what 2>happen 3>on 4>otv 5>on 6>friday 7>night 8>will 9>make 10>most 11>people 12>exclaim 13>in 14>exasperation 15>: 16>that 17>be 18>enough 19>!

0-0 poate /maybe-0

1-16 ca /that-16

2-2 intimplare /happen-2

3-3 de /on-3/on-5/in-13

4-6 vineri /friday-6

5-7 seara /night-7

…

Note that the up-bottom process, illustrated above, initially aligned the Romanian position 3 with the English position 3, because this one was consecutive with 2. The precedence constraint proved this alignment to be wrong and corrected it from 3 to 5.

4.3. *Pair alignments*

This linguistic assumption applies at the level of finding alignments which translation equivalence dictionary says nothing about. It consists in taking pairs of parts-of-speech depending, in some extent, on each other. For instance, if there are sequences adjective-noun (irrespective of their mutual order) both in Romanian and in English and nouns are paired together, then it is very likely to associate the adjectives too. So for example, in the parallel text below such an adjective-noun sequence is given by the Romanian positions 0 and 1 (i.e.

chinuit arestare) and by the English positions 2 and 3 (*lame arrest*). TREQ only found the equivalence *arestare-arrest* and aligned the Romanian position 1 with English position 3. The assumption discussed here allows to put in correspondence the adjectives *chinuit* (at Romanian position 0) and *lame* (at English position 2) and creates, therefore, the new correct alignment 0-2.

Exp. 7 **RO:** 0>chinuit 1>arestare 2>al 3>lui 4>treptow
EN: 0>treptow 1>'s 2>lame 3>arrest

1-3* arestare /arrest-3*
4-0* treptow /treptow-0*

The same holds for other kinds of pairs, such as noun-noun, preposition-noun/adjective/ numeral, conjunction-verb/adverb, article-noun/numeral etc.

4.4. *Dictionary collocations*

In some cases, TREQ associates a word in a language with two words in the other language that happen to form a collocation. For instance, one can find, as translation equivalences in the TREQ lexicon, both *liceu=school* and *liceu=high*, while *high school* is an English collocation, and the proper translation equivalence is *liceu=high school*. In order to use this fact in word alignment, we assume that if two words in a text are consecutive and both are translation equivalents of the same word in the other language, then both of them should be aligned with the one. Thus, it results a 2:1 alignment. Consider the following example.

Exp. 8 **Dictionary entries**:
post,N tv,N 146.58
post,N station,N 280.97
Bilingual text:
0>post 1>acesta 2>pentru 3>care 4>sorin 5>rosca …
0>this 1>tv 2>station 3>for 4>which 5>" …
Alignment map:
0-1 post /tv-1/station-2
1-0 acesta /this-0/a-20
2-3 pentru /for-3
3-4 care /which-4
4-11 sorin /sorin-11
5-12 rosca /rosca-12

One can see that the pairs *post-tv* and *post-station* have got good scores (the figures on the right) for counting as valid translation equivalences in the dictionary. On the other hand, in the bilingual text to be aligned, both English words are adjacent. This entitles the assumption that the two English words translate together the Romanian one and therefore the right alignment is Romanian:0- English: 1 2.

4.5. *Language-specific rules*

Besides these assumptions we have applied some language-specific rules concerning Romanian versus English syntax particularities or cross-linguistic differences in part-of-speech mapping. Without getting into details, we only mention some few examples. Usually, the English phrases of two nouns: noun1 noun2, e.g. *chocolate candy*, are translated into Romanian as noun2-the preposition 'de'-noun1: *bomboană de ciocolată*. On the other hand, Romanian has a lot of articles and particles mapping into English determiners and prepositions, respectively. The language-specific module is a distinct unit in TREQ-AL, in order to keep the

generality of the algorithm, and this module can be adapted for other pairs of languages than Romanian and English.

5. Conclusions

As we have already said, the applying of the linguistic methods discussed above managed to improve the results of the algorithm TREQ-AL from the values obtained at the moment of the shared task: P=81.29%, R=60.26 and F=69.21, to significantly better ones at the moment: P=84.95, R=65.13, F=73.73. In order to get a graphical illustration about the improvement contribution of each method, we have done the following evaluation experiment. We have disabled each method in turn and we have noticed the new results of the algorithm with respect to the considered measures. The data presented in the table below show how much a measure grows or decreases by disabling the respective method.

Linguistic method	Precision [%]	Recall [%]	F-measure [%]
Cognate detection	-0.42	-0.64	-0.57
Precedence constraint	-1.92	-0.43	-1.00
Pair alignments	+0.01	-0.30	-0.19
Collocations	+0.15	-0.22	-0.09
Language specific rules	+1.07	-3.14	-1.68

Table 1. Method percentage contribution

As one can see, the language specific rules bring the most important contribution to the f-measure value. By deactivating this module the precision grows indeed but the recall dramatically decreases and the f-measure, either. It turns out that this is a necessary module and, as a general conclusion, that a language-oriented algorithm could be better than a general one. The next important share (of 1%) is due to the precedence constraint. This method increases the precision with almost 2% and the recall with about 0.5%, while the cognate detection improves both the precision and the recall with about half a percent.

We hope these data offer enough evidence about how much some simple linguistic assumptions can contribute to the general results of a statistical analysis on parallel texts.

References

Brants T. (2000). TnT – A statistical Part-of-Speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference*, ANLP-2000, April 29-May 3, Seattle, WA.

Dejean H., Gaussier E., Goutte C. and Yamanda K. (2003). Reducing Parameter Space for Word Alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, May-June, Edmonton, Canada: 23-26.

Gale W.A. and Church K.W. (1993). A program for Aligning Sentences in Bilingual Corpus. *Computational Linguistics*, vol. (19/1): 75-102.

Melamed D. (2000). Models of translation equivalence among words. *Computational Linguistics*, vol. (26/2): 221-249.

Mihalcea R. and Pedersen T. (2003). An Evaluation Exercise for Word Alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, May-June, Edmonton, Canada: 1-10.

Tufis D. and Barbu A.-M. (2002). Revealing Translators' Knowledge: Statistical Methods in Constructing Practical Translation Lexicons for Language and Speech Processing. *International Journal of Speech Technology*, vol. (5): 199-209.

Tufis D., Barbu A.-M. and Ion R. (2003). TREQ-AL: A word alignement system with limited language resources. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, May-June, Edmonton, Canada: 36-39.

Zhao B. and Vogel S. (2003). Word Alignment Based on Bilingual Bracketing. In *Proceedings of the HLT-NAACL 2003* Workshop on *Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, May-June, Edmonton, Canada: 15-18.

Gli sbarchi dei clandestini nei quotidiani: un'analisi testuale esplorativa

Silvia Bartoletti, Alessandra Garbero, Silvia Montecolle,
Ferdinando Nisco, Emanuela Recchini, Irene Salerno

Università di Roma "La Sapienza", Facoltà di Economia e Commercio
silbartoletti@hotmail.com, Alessandra.Garbero@fao.org, fmontecolle@tiscali.it,
nisco@freemail.it, emanuelarecchini@libero.it, irenesalerno@yahoo.it

Abstract

In this paper we aim to present a method of textual analysis providing a brief description of immigration through the articles published in different newspapers in spring 2003 when Sicily – Italy – was invaded by a great number of clandestines. We have considered five newspapers representing five different Italian geographic areas (north-east, north-west, centre, south and islands). First of all, we aim to analyse forms and verbs used, highlighting specific features and analogies; then, using automatic techniques we aim to have a synthetic knowledge of different points of view picking up the themes of the texts, and the positions that newspapers hold by means of "forms-texts" tables.

Riassunto

Il presente lavoro ha al centro del suo interesse l'analisi del fenomeno migratorio per come esso è stato rappresentato e descritto in alcuni quotidiani italiani nel periodo della primavera dell'anno 2003. La scelta del "momento storico" preso in esame non è stata casuale: l'Italia, infatti, è stata interessata, proprio in tale periodo, da un'ondata migratoria che si è imposta sulle cronache dei principali quotidiani italiani per via della particolare drammaticità con cui si sono svolti gli sbarchi di stranieri clandestini. Analizzando con gli strumenti propri dell'analisi testuale gli articoli estratti da cinque diverse testate giornalistiche italiane scelte secondo un criterio "geografico" ci si è proposti di valutare in che modo i quotidiani scelti hanno affrontato il problema migratorio, cogliendo analogie e specificità attraverso un'analisi multidimensionale dei dati testuali.

Parole chiave: analisi delle corrispondenze semplici, analisi testuale, flussi migratori.

1. Introduzione

Nel periodo tra maggio e giugno 2003, la buona stagione e le favorevoli condizioni del mare hanno contribuito all'improvvisa ripresa degli sbarchi sulle coste meridionali dell'Italia di imbarcazioni cariche di clandestini. In particolare, è stata l'isola di Lampedusa, in Sicilia, la principale meta degli sbarchi. I quotidiani italiani hanno discusso per più di un mese dell' argomento nelle pagine dedicate alla cronaca.

Sono stati raccolti articoli omogenei per contenuto – cronaca – e dimensioni, apparsi a cavallo tra i mesi di maggio e di giugno sui seguenti quotidiani: Il Gazzettino Veneto, Brescia Oggi, Il Messaggero, Il Mattino, La Sicilia.

Il criterio con il quale è stata operata la scelta delle testate ha seguito un disegno preciso: si è scelto di adottare un parametro "geografico", per aver così rappresentata, attraverso quotidiani locali, l'Italia tutta intera: il Nord Est (Il Gazzettino Veneto), il Nord Ovest (Brescia Oggi), il

Centro (Il Messaggero), il Sud (Il Mattino); per quanto riguarda, poi, il Meridione, si è ritenuto interessante prendere in considerazione anche il quotidiano "La Sicilia", per via del peculiare rapporto che questa regione ha intrattenuto con gli sbarchi di stranieri, nella scorsa primavera.

Obiettivo di questa ricerca è stato quello di esplorare, attraverso un'analisi testuale, le parole-chiave caratterizzanti il corpus *in toto,* e le specificità tra i quotidiani oggetto d'analisi, per cogliere similitudini o disuguaglianze dettate da ragioni geografiche – e forse politiche – nel pensiero e nella percezione dei giornali esaminati.

2. Il corpus e l'analisi delle forme grafiche

Il corpus risulta composto da circa 37.000 occorrenze totali (N), e da circa 6.380 forme grafiche diverse (V), equidistribuite tra i quotidiani oggetto di studio.

Forme grafiche	Occorrenze	Forme grafiche	Occorrenze	Forme grafiche	Occorrenze
Clandestini	213	isola	85	donne	59
Lampedusa	175	Isola	85	ministro	56
Immigrati	142	bordo	84	costiera	54
Mare	107	persone	84	barcone	52
Italia	100	Libia	83	notte	50
immigrazione	97	Bossi	60	legge	47
guardia	92	porto	60	Lega	38

Tabella 1. Graduatorie delle forme grafiche maggiormente impiegate nel corpus – valori assoluti

In primo luogo si è proceduto con lo studio delle forme grafiche in termini di frequenza d'uso, eliminando gli *hapax* e le parole strumentali quali preposizioni, articoli, ecc. Nella graduatoria delle forme grafiche, costruita rispetto al numero di occorrenze, (tabella 1) la prima parola piena che si incontra è "clandestini", (che nel *corpus* presenta un numero di occorrenze pari a 213), forma grafica a connotazione prevalentemente negativa, che presenta un numero di occorrenze superiore rispetto ad "immigrati" ed "immigrazione", a pregnanza semantica prevalentemente neutra. Seguono forme grafiche che individuano geograficamente i luoghi dove sono avvenuti gli sbarchi: "Lampedusa", "mare", "Italia" ed "isola". Anche "persone" e "bordo" sono parole chiave per il *corpus*: la prima indica in maniera ancora più generale i soggetti coinvolti nel fenomeno; la seconda rappresenta il modo in cui queste persone arrivano in Italia: a "bordo" di qualche mezzo, che all'interno del *corpus* è legato a diversi termini (barcone, barca, peschereccio, imbarcazione, ecc.).

Analisi delle forme verbali

Di particolare interesse si è rilevato lo studio delle forme verbali.

Si è proceduto con la lemmatizzazione (riconducendo tutte le voci verbali al corrispondente tempo all'infinito) riducendo il numero di forme da esaminare di circa il 50%, passando dai precedenti 1.338 differenti tempi verbali a 692 voci all'infinito.

Escludendo i verbi ausiliari e tutti i servili[1], la graduatoria dei verbi all'infinito col maggior

[1] Tali verbi hanno fatto riscontrare una prevalenza nettamente superiore a tutti i rimanenti. Tale situazione non permetteva di far emergere i reali aspetti caratteristici del testo. Si è pensato, quindi, di escluderli da alcune analisi.

numero di voci coniugate nel *corpus* vede al primo posto "venire" con 60 occorrenze, seguito da "trovare" con 45, da "dire" con 43, da "arrivare" con 40 e da "chiedere" con 37 (tabella 2). Ad un elevato numero di occorrenze non è corrisposto un elevato utilizzo di differenti unità lessicali, ossia di diverse forme verbali coniugate. Nel caso specifico, è risultato che "chiedere", benché all'ultimo posto della graduatoria delle occorrenze, è al secondo posto in quella delle unità lessicali.

Verbi	Occorrenze	Unità lessicali
Venire	60	13
Trovare	45	7
Dire	43	9
Arrivare	40	10
Chiedere	37	10

Tabella 2. Graduatoria dei verbi maggiormente impiegati nel corpus – valori assoluti

Per individuare i verbi maggiormente caratterizzanti il corpus si è effettuato un confronto fra questo ed un modello di riferimento[2]. In particolare, calcolando uno scarto sulle occorrenze si sono individuati i verbi sovra-utilizzati rispetto allo standard scelto: le cosiddette "parole chiave" del *corpus*. Come era logico supporre, i valori più elevati sono stati registrati per verbi strettamente correlati con la descrizione della tematica in esame (tabella 3).

Verbi	Scarto	Verbi	Scarto
Avvistare	98,15	Scortare	27,61
Spiaggiare	58,35	Salpare	23,56
Trasbordare	31,95	Approdare	17,19
Naufragare	29,40	Sbarcare	16,66
Intercettare	27,84	Speronare	16,60

Tabella 3. Graduatoria dei primi 10 verbi con scarto positivo più elevato

Considerando la frequenza cumulata associata ai primi 10 verbi maggiormente caratterizzanti il corpus[3] si è osservato che ne La Sicilia e ne Il Gazzettino Veneto è stato registrato il numero di occorrenze più elevate, rispettivamente pari a 44 e 35. Tale situazione sancisce l'impostazione maggiormente indirizzata al mero racconto dei fatti di cronaca da parte di queste due testate rispetto alle rimanenti.

3. L'analisi delle specificità nei sub-testi

La tecnica delle specificità permette di comparare gli articoli provenienti dalle differenti testate e di osservare le modificazioni nel profilo lessico-metrico. In sintesi le specificità rappresentano il sovra-impiego o viceversa il sotto-impiego di una forma rispetto ad una soglia di probabilità che nel nostro caso è posta pari al 5%.

Il quotidiano Brescia Oggi rispetto alle altre testate in esame ha presentato un sovra-utilizzo di termini più legati alla una discussione politica del problema. La maggior parte delle forme

[2] Il modello di riferimento scelto è rappresentato dagli articoli apparsi nel corso del 1990 sul giornale La Repubblica, a tiratura nazionale.

[3] Sono tali i verbi con scarto positivo più elevato.

grafiche sono di natura politica (“Bossi”, “Lega”, “Pisanu”, “maggioranza”, “accordo” “presidente”, “Berlusconi”). Il test statistico ha mostrato come, la forma grafica “Bossi” risulti altamente specifica in questo giornale, con una probabilità infinitamente nulla (+E12) di essere un evento puramente casuale. Ha una specificità positiva anche “immigrazione”, termine che risulta negli articoli di Brescia Oggi molto più utilizzato di “immigrati” o “clandestini”. Questo induce a sostenere che in Brescia Oggi sia stata data maggiore attenzione alla questione immigrazione e alla ricerca di una soluzione politica, piuttosto che alla semplice descrizione degli eventi che hanno visto protagonisti i clandestini. Tale osservazione è avvalorata anche dall’analisi dei termini sotto-utilizzati (“guardia”, “porto”, “isole”, “donne”, “costiera”), forme in genere utilizzate per descrivere la notizia.

Un taglio decisamente diverso hanno gli articoli presenti ne Il Gazzettino Veneto. In questo caso le forme grafiche maggiormente utilizzate sono quelle che consentono di descrivere l’accaduto con toni drammatici (“cadaveri”, “disperati”, “bilancio”, “carretta”…), mentre sono sotto-utilizzati i termini politici.

Il Messaggero presenta nei suoi articoli un numero inferiore di forme grafiche che riguardano gli aspetti legislativi. Le specificità non confermano però del tutto le ipotesi avanzate sulla base delle occorrenze: Il Messaggero, infatti, non è caratterizzato da un gran numero di termini relativi alla cronaca.

L’analisi delle specificità evidenzia negli articoli de Il Mattino forme che affrontano la questione in termini di soluzioni da ricercare per il futuro: “intervento”, “Paese”, “futuro”, “legge” potrebbero far emergere la necessità della ricerca di una potenziale soluzione. Sono sotto-utilizzate alcune forme che nell’analisi sono state sempre associate alla descrizione del fatto di cronaca.

La Sicilia, in quanto giornale locale della regione più direttamente coinvolta nell’evento “sbarchi”, utilizza nei suoi articoli maggiori riferimenti alle persone coinvolte (“extracomunitari”, “clandestini”, “donne”, “bambini”), ai luoghi dove gli sbarchi sono avvenuti (“isola”, “Lampedusa”), alle forze dell’ordine preposte al controllo dei mari per limitare gli incidenti (“guardia”, “motovedette”, “Capitaneria”, “polizia”, “costiera”, “finanza”) e al tipo di organizzazione impiantata per accogliere gli immigrati (“organizzazione”, “struttura”, “ospedale”). D’altra parte sono sotto-utilizzate le forme politiche.

4. Analisi dei segmenti ripetuti

Le analisi quantitative condotte sui testi presuppongono una suddivisione del *corpus* in unità elementari generando una forte decontestualizzazione, poiché tale operazione annulla il tessuto dei legami sintagmici tra le parole.

Utilizzando le forme grafiche ci si rende conto che, venendo a mancare del tutto il riferimento al contesto, è difficile stabilire in quali temi specifici si attualizza il senso generico di ciascuna parola. I segmenti ripetuti sono le unità statistiche che meglio si prestano ad assolvere questo compito analitico, in quanto costituiti da sequenze di forme grafiche che si ripetono identiche in un testo: questa loro natura li rende particolarmente adatti ad evidenziare i legami sintagmatici stabili tra gruppi di forme.

Nell’ambito dello studio del *corpus* oggetto d’analisi attraverso l’osservazione dei segmenti ripetuti, ci si propone di definire se esiste una strutturazione tematica e stilistica condivisa di trattare il tema migratorio indipendentemente dal quotidiano.

Forme grafiche	Brescia Oggi	Gazzettino Veneto	Il Messaggero	Il Mattino	La Sicilia
Barca			13 +E02	1 -E03	
Barcone				4 -E02	21 +E03
Berlusconi	7 +E03				
bordo				30 +E02	
Bossi	28 +E12		0 -E07		
cadaveri		12 +E03			
Capitaneria					10 +E03
chiede					0 -E02
clandestini				19 -E06	77 +E06
disperati		19 +E03			
donne	22 +E03				
extracomunitari					24 +E07
figli			8 +E04		
guardia	1 -E04				
immigrati	7 -E03				
immigrazione	32 +E09				
intervento				10 +E04	
isola				8 -E03	41 +E07
Italia					14 -E02
Lega	18 +E08		0 -E04		
legge			0 -E05	19 +E03	
maggioranza	6 +E04				0 -E02
ministro	23 +E09				
notte			19 +E03	3 -E03	
organizzazione					12 +E07
Paese				8 +E03	
paesi			1 -E02		
peschereccio			11 +E03		
Pisanu	17 +E07		0 -E05-		
polizia					12 +E03
presidente	10 +E04				
problema		13 +E02			
terra			11 +E03		
vittime			6 +E03		

Tabella 4. Specificità positive e negative nei quotidiani

Come per le forme grafiche, anche per i segmenti è possibile fissare delle soglie di frequenza che ne condizionano la selezione nel testo (frequenza maggiore o uguale a 4); la soglia scelta è abbastanza bassa per ottenere una lista esaustiva. Ma l'elenco ottenuto, così com'è, non è efficiente: risultano selezionate molte sequenze non significative perché costituite soltanto da forme grammaticali, incomplete o semplicemente banali ai fini dell'analisi. Sono stati eliminati questi segmenti e si è proseguita la selezione ricorrendo al criterio di "completezza grammaticale" (Morrone, 1993) per limitare drasticamente il problema della ridondanza.

L'analisi dei segmenti ripetuti è stata effettuata sulla base della lista estratta, che prevedeva un elenco totale di 1.520 segmenti. Tra questi sono stati presi in esame quelli che presentavano un valore dell'Indice di Significatività superiore a 0,90 (tabella 5).

Segmento	Occorrenze totali	Lunghezza	Indice IS	Indice IS relativo
centro di accoglienza	54	3	2,85	0,31
navi della marina militare	10	4	2,84	0,17
presunti scafisti	11	2	2,73	0,68
carretta del mare	29	3	2,59	0,28
peschereccio tunisino	8	2	2,25	0,56
acque internazionali	16	2	2,14	0,53
tratto in salvo	7	3	2,02	0,22
motovedetta della guardia	18	3	1,69	0,18
vita migliore	7	2	1,52	0,38
nostro Paese	6	2	1,50	0,37
barca da pesca	7	3	1,47	0,16
bilancio ufficiale	4	2	1,41	0,35
coste siciliane	6	2	1,36	0,34
nostre coste	7	2	1,26	0,31
immigrazione clandestina	8	2	1,09	0,27
cadaveri recuperati	5	2	1,05	0,26
nord africani	6	2	1,02	0,25
verso l'Italia	9	2	0,94	0,23

Tabella 5. Segmenti ripetuti con IS>0,90

Gli articoli presi in esame, nel sottolineare come il <nostro Paese> rappresenti per alcuni popoli un'attrazione fortissima, non dimenticano di rimarcare la precarietà delle condizioni di viaggio degli extracomunitari (viaggiano trasportati da <presunti scafisti> a bordo di quelle che vengono definite le <carrette del mare>): molto spesso, difatti, la "rotta della speranza" Nordafrica-Sicilia si trasforma in tragedia. Naufragano <barche da pesca> in pessime condizioni che inseguono il sogno di una <vita migliore>, trasformato in incubo di morte (numerosi i <cadaveri recuperati>). Solo l'intervento delle <navi della Marina Militare>, affiancate dalle <motovedette della Guardia Costiera>, evita l'irreparabile. Non sempre la politica riesce a gestire la situazione; accanto ad organismi dello stato, (<Marina Militare>, <Guardia Costiera>, <Guardia di Finanza>), ad affrontare l'emergenza vi sono i <centri di accoglienza> ed organizzazioni di volontariato laiche e confessionali.

5. Analisi delle corrispondenze semplici

Per analizzare al meglio le relazioni multidimensionali tra i quotidiani oggetto dello studio e le forme grafiche utilizzate da ciascun giornale, è stata realizzata un'analisi delle corrispondenze semplici con il software statistico SPAD che, sintetizzando le informazioni contenute in ciascun subtesto, evidenzia associazioni e contrasti, al fine di cogliere al meglio analogie e specificità di ciascun quotidiano.

Il *corpus* è stato preventivamente trattato, eliminando le congiunzioni, i pronomi, le preposizioni e le esclamazioni. Si è ritenuto proficuo concentrare l'analisi sui primi due fattori che insieme sintetizzano una discreta quota di inerzia complessiva (tabella 6).

Fattore	Autovalore	Percentuale	Percentuale cumulata
Primo	0,1839	38,18	38,18
Secondo	0,1286	26,71	64,89
Terzo	0,0999	20,74	85,63
Quarto	0,0692	14,37	100,00

Tabella 6. Autovalori ed inerzia spiegata dai primi quattro fattori

L'analisi ha fornito anche le coordinate e i contributi sugli assi fattoriali delle cosiddette frequenze attive che, nel caso specifico, corrispondono ai cinque sub-testi considerati.

Le coordinate permettono di individuare la posizione dei cinque quotidiani sul piano fattoriale (Il Gazzettino Veneto è quello più vicino al baricentro). Proseguendo con l'analisi dei contributi delle frequenze attive sui primi quattro fattori, quello che si evince è che sul primo asse Brescia Oggi è il quotidiano che più apporta un contributo significativo (48%), seguito da Il Mattino (23%). Sul secondo asse, invece, i quotidiani che apportano il contributo maggiore sono Il Messaggero (61.4%) e La Sicilia (35%); quasi nullo il contributo de Il Gazzettino Veneto ai primi due assi (rispettivamente 1,2% e 2,2%).

Analizzando i contributi degli individui attivi sui primi quattro fattori, ad influire di più sul primo asse sono forme grafiche di tipo politico-istituzionale ("Pisanu", "Bossi", "presidente", "Paesi", "uomo", "legge Bossi-Fini", "Berlusconi", "UE", "cooperazione", "maggioranza", "chiede", "Senato", "Camera", "vertice", "voce", "Calderoni", "Frattini", "Carroccio").

Sul secondo asse, invece, hanno peso maggiore le forme grafiche di tipo descrittivo (Lampedusa, mare, Libia, guardia di finanza, motovedette, problema, paesi, qui, peschereccio, terra, ospedale, organizzazione, casa, Somalia, racket) e discorsivo ("racconta", e tutte le forme verbali alla terza persona singolare e plurale).

Il primo asse rappresenta la tipologia di informazione che viene fornita dal quotidiano: in particolare "politica" (semiasse negativo) e "non politica" (di cronaca, pertanto) (semiasse positivo). Il secondo asse, invece, è indicatore della dimensione territoriale dell'informazione: nazionale (semiasse positivo), locale (semiasse negativo).

In figura 1 viene proposto il grafico con la rappresentazione simultanea delle frequenze e degli individui attivi sui primi due assi fattoriali: ciascun giornale ha una caratterizzazione ed una collocazione a sé stante.

Nel primo quadrante (++) è posizionato Il Messaggero. Quella de Il Messaggero è cronaca nazionale "tout court", lontana dalla politica e dalle problematiche locali e regionali, ma vicina all'Italia tutta intera; è cronaca nel senso più vero della parola: descrittiva, innanzitutto (pescherecci, acque, barche, uomini, frontiera, terra, notte, assistenza), anche se talvolta assume tinte patetiche (cadaveri, casa, tragedia, vittime) o polemiche (racket).

Nel secondo quadrante (-+) è posizionato Brescia Oggi; nonostante non sia un giornale di partito, ciò che emerge è la sua natura politica (specie se paragonato agli altri quotidiani oggetto di studio), probabilmente per la realtà geografica di cui è portavoce (Brescia e zone limitrofe) che lo rende inequivocabilmente vicino a Bossi, affrontando il problema in chiave nazional-politico-legislativo (Bossi, senato, maggioranza, presidente, Pisanu, legge Bossi-Fini, immigrazione, Berlusconi, pattugliamento, problema, emergenza).

Nel terzo quadrante (--) è posizionato Il Mattino. Analogamente a Brescia Oggi si sofferma sulla dimensione legislativa del problema, anche se in chiave meno politica (legge Bossi-Fini, fenomeno, Italia, acque internazionali); sul secondo asse non è distante da La Sicilia, probabilmente per la sua natura di "quotidiano del Sud" (asilo, accoglienza, forze, ordine, intervento).

Nel quarto quadrante (+-) sono posizionati La Sicilia ed Il Gazzettino. La Sicilia è completamente assorbita dalla dimensione locale del problema. In primo piano tutta una serie di specificità geografiche (Lampedusa, Agrigento, Palermo, Porto Empedocle) ed un resoconto puntuale di tutti gli aspetti organizzativi ed assistenziali (organizzazione, allarme, ospedale, guardia costiera, capitaneria, motovedette, guardia di finanza).

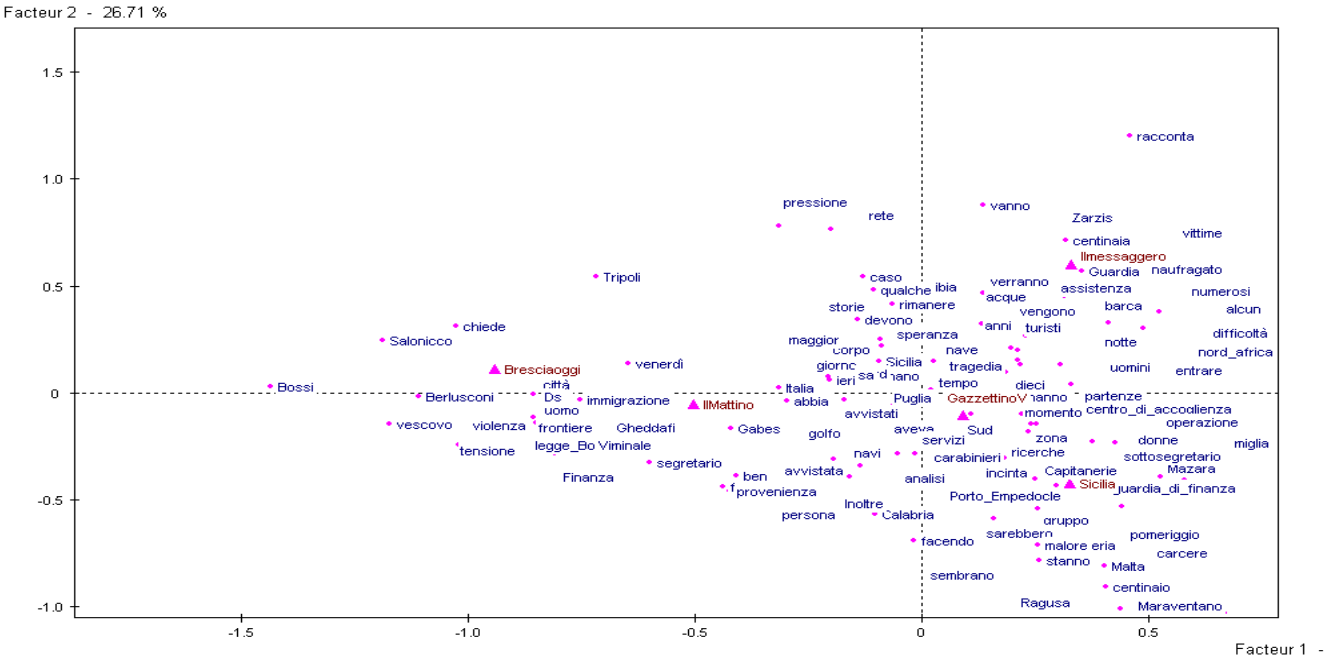


Figura 1.Diagramma quotidiani-forme grafiche

6. Conclusioni

L'analisi è stata condotta su un argomento che durante l'ultima primavera ha suscitato particolare interesse, sia per le problematiche organizzative, che un tale flusso di immigrati ha comportato, sia per l'approccio legislativo, che in qualche modo doveva far fronte alla sopraggiunta emergenza, sia per l'aspetto tragico dei numerosi naufragi avvenuti.

L'analisi testuale degli articoli dei cinque differenti giornali raccolti ha portato ad interessanti conclusioni, facendo emergere le differenti impostazioni delle diverse testate.

In particolare, si osservata la netta contrapposizione tra Brescia Oggi, che ha affrontato la questione facendo maggiormente riferimento a spunti legislativi e La Sicilia, che in quanto giornale dell'isola direttamente coinvolta dagli sbarchi, è stata più attenta a descrivere come si è fatto fronte all'emergenza. Su posizioni intermedie si sono attestate le rimanenti tre testate, pur denotando un'impostazione più politica per il Mattino e un'impostazione più di cronaca per Il Messaggero e Il Gazzettino Veneto.

Bibliografia

Bolasco S. (1999). *Analisi Multidimensionale dei dati*. Carocci.

Bolasco S. e Cipriani R. (1995). *Ricerca qualitativa e computer*. Franco Angeli.

Lebart L., Salem A. e Berry L. (1998). *Exploring textual data*. Kluwer Academic Publishers.

Lebart L., Morineau A. e Piron M. (1995). *Statistique Exploratoire Multidimensionnelle*. Dunod.

Lebart L. e Salem A (1994). *Statistique textuelle*. Dunod.

Morrone A. (1993). Alcuni criteri di valutazione della significatività dei segmenti ripetuti. In *JADT 1993*.

Contribution de la métrique à la stylométrie

Valérie Beaudouin[1], François Yvon[2]

[1]FT R&D – 38-40 rue du Général Leclerc – Issy les Moulineaux– France
valerie.beaudouin@francetelecom.com

[2]GET/ENST et CNRS/LTCI – 4, rue Barrault – 75013 Paris – France
francois.yvon@free.fr

Abstract

In this paper, we present the results of our first attempts to use metrical informations for authorship attribution. 58 plays in verse by Corneille, Racine and Molière were systematically analyzed regarding syllabic, morpho-syntactic and stress structure, using the Metrometer, a Natural Language Processing tool aimed at the analysis of French classical verse. Our main finding are that (i) Corneille and Racine are very consistent in their metrical use of words, Molière being quite less so; (ii) using statistical language modeling tools, it is in fact possible to build syllable-based model which can effectively discriminate genre and authors.

Résumé

Nous montrons comment des informations métriques du vers peuvent être utilisées dans le cadre d'études stylométriques. L'ensemble des pièces en vers de Corneille, Racine et Molière, soit 58 pièces, a été analysé par le Métromètre, un outil d'analyse du vers classique qui produit pour chaque position métrique un ensemble de descriptifs linguistiques (syllabe, catégorie morpho-syntaxique, accent…). Alors que Corneille et Racine sont cohérents à travers leur œuvre dans leur manière de procéder aux décomptes métriques dans le vers, Molière est quant à lui plus fluctuant, au moins sur trois mots. Toutefois, lorsqu'on les considère de manière indépendante, les différents descripteurs des positions métriques utilisés ici ne permettent pas de différencier de manière sûre des auteurs pour un genre donné. En revanche, l'utilisation de modèles statistiques du vers, fondés sur la séquence syllabique, permet de construire des outils de discrimination qui conduisent à identifier avec une précision raisonnable genres et auteurs.

Mots-clés : stylométrie, modèles de langage.

1. Introduction

Les questions d'attribution des textes ont régulièrement suscité des débats passionnels, comme les travaux anglo-saxons sur les textes de Shakespeare ou les récentes études sur Corneille et Racine (Labbé et Labbé, 2001). Les polémiques s'intensifient avec la notoriété des auteurs mis en jeu : changer l'attribution de pièces classiques revient à mettre en péril des savoirs qui s'appuient souvent sur les biographies des auteurs pour comprendre les œuvres. L'intérêt de la sphère médiatique pour ces remises en cause tend à aggraver les polémiques. Les travaux de stylométrie qui traitent des questions d'attribution dérangent car ils mobilisent des techniques encore peu répandues dans le milieu littéraire et souvent font abstraction de toute la connaissance accumulée dans le domaine, passent parfois rapidement sur des notions aussi centrales que celle de genre. Les débats sont également intenses entre les chercheurs du domaine, comme en témoigne la synthèse qu'Holmes (1998) propose des travaux en stylométrie.

Différents types d'indicateurs ont été utilisés dans les travaux d'attribution (mots les plus fréquents, mots outils, rimes….) (Holmes, 1998). Nous proposons de tester les éléments métri-

ques (syllabes, catégories morpho-syntaxiques et accents) comme candidats potentiels à une différenciation des auteurs et des genres. Y a-t-il une manière propre à chaque auteur de faire des vers ? En quoi les aspects métriques peuvent-ils qualifier l'écriture ? Nous nous appuyons sur les pièces en vers de Corneille, Molière et Racine[1] en accordant une place particulière aux comédies, puisque les différences de genre l'emportent généralement sur les écarts entre auteurs. Les aspects métriques sont traités avec un outil, le métromètre, mis au point en 1993 pour l'analyse du vers classique (Beaudouin et Yvon, 1995).

Après une brève présentation du corpus utilisé et du métromètre, nous montrons que le traitement métrique de certains mots dans le vers peut permettre de distinguer les auteurs entre eux. Ces indices très ténus doivent être complétés par une approche plus globale du vers. Après avoir montré que des statistiques descriptives sur la plupart des critères métriques ne permettent pas de distinguer auteurs et genres, nous recourons aux modèles de langage appliqués aux séquences de syllabes métriques. Des modèles sont construits pour les auteurs et les genres, sur des échantillons de pièces, et on teste la capacité du système à attribuer les vers restant au bon modèle.

2. Corpus et métromètre

Le corpus est constitué par l'ensemble des pièces en vers de Corneille, Molière et Racine, soit 58 pièces.

	Comédies	Tragédies	Pièces diverses
Corneille	9	21	4
Racine	1	11	–
Molière	12	–	–

Tableau 1. Répartition des pièces du corpus par auteur et genre

Les éditions électroniques utilisées correspondent à l'édition de Marty-Laveaux (1862) pour Corneille – qui intègre toutes les corrections faites en 1660 par Corneille sur ses premières pièces, à celle de Paul Mesnard (1885) pour Racine et à celle d'Eugène Despois (1873) pour Molière[2].

Le métromètre est un outil de description systématique de la structure phonétique, morpho-syntaxique et accentuelle du vers. Il repose sur une analyse phonétique et métrique du vers. Cet outil résulte de l'adaptation au cas particulier du vers d'un phonétiseur du français développé par François Yvon (1995). Ce phonétiseur d'appuie lui-même sur un analyseur syntaxique, Sylex, développé par Patrick Constant (1991).

Le métromètre, après analyse syntaxique, transcrit le vers dans l'alphabet phonétique, le découpe selon les positions métriques, en respectant les règles de la versification (diérèse/synérèse, décompte du *e* muet et liaison) et finalement, attribue à chacune des positions un certain nombre de marquages ou étiquettes d'ordre exclusivement linguistique (syllabe et

[1] Les œuvres de Corneille et Racine ont été les premières à être analysées en France avec les méthodes de statistique lexicale, développées par Charles Muller (Muller, 1967 ; Bernet, 1983). Ces travaux de lexicométrie ont permis entre autre de mettre en évidence les spécificités lexicales liées aux genres, aux auteurs et aux périodes.

[2] Le corpus Molière nous a été fourni par Charles Bernet, que nous remercions. Cette édition électronique des pièces a été mise en place dans le cadre d'un projet INALF sur le théâtre du XVII^{e} : les pièces y sont balisées dans un langage apparenté au XML, ce qui permet une préparation accélérée pour le passage au métromètre.

voyelle métriques, fin de mot, catégorie morpho-syntaxique, accent). Voici ce que donne l'analyse du vers suivant de Racine :

De cette nuit, Phénice, as-tu vu la splendeur ?

Racine, *Bérénice*, vers 302.

Syllabes métriques	d ə	s ɛ	tɛʁ	n ɥi	f e	n i s	a	t y	v y	l a	spl ɑ̃	d œʁ
Voyelles métriques	ə	ɛ	əʁ	ɥi	e	i	a	y	y	a	ɑ̃	œ
Repérage des fins de mots (fdm)	fdm	-	fdm	fdm	-	fdm	fdm	fdm	fdm	fdm	-	fdm
Catégories syntaxiques	Préposition	pronom, dét	pronom, dét	nom	nom propre	nom propre	verbe	pronom, dét	adjectif	pronom, dét	nom	nom
Marquage accentuel (accent)	-	-	-	accent	-	accent	-	accent	accent	-	-	accent
Nb syllabes	12											

Les productions du métromètre mises sous forme de base de données peuvent être explorées dans plusieurs directions : construction de la figure globale du vers (phonétique, syntaxique, accentuelle) par pièce, genre, auteur… ; recherche de vers répondant à certaines contraintes (vers constitués de très peu de mots, avec peu ou beaucoup de variations phonétiques…) ; recherche de corrélations entre marquages (liens entre le contenu lexical et la forme métrique, entre le genre et la rime…). Nous proposons une nouvelle exploitation de la base de vers enrichie de traits de description pour explorer les questions d'attribution.

3. Les contours de la syllabe comme marque d'auteur

L'application du métromètre à l'ensemble des pièces permet d'identifier des variations dans le traitement métrique. Nous avions déjà observé pour Corneille comme pour Racine une très grande cohérence dans la définition des syllabes métriques. Un mot donné était traité de la première à la dernière pièce de la même manière. Il faut rappeler que Corneille a procédé en 1660 à une révision de ses premières pièces. Dans les trente années qui séparent sa première pièce de 1660, la prosodie a connu des évolutions notables, qui ont été prises en compte dans ses révisions. Celles-ci contribuent à donner le sentiment d'une grande cohérence métrique de bout en bout, quel que soit le genre. Par ailleurs, nous avions noté une seule différence entre Corneille et Racine liée au traitement métrique de *hier* : pour Corneille, il constitue une syllabe, tandis qu'il en représente deux pour Racine. Comment se situe Molière par rapport au traitement de la syllabe métrique ?

Pour ce faire, nous avons sélectionné tous les vers qui d'après le métromètre étaient constitués de 11 ou 13 syllabes. Ces vers constituent forcément des erreurs ou variations de versification puisqu'il n'y a aucun vers de ces longueurs là dans le corpus. Cette sélection permet d'identifier les mots dont l'analyse phonético-métrique se distingue de celle de Corneille et Racine, ces auteurs ayant servi à étalonner le métromètre. Ensuite, sont extraits tous les vers qui contiennent ces mots. Ainsi, peuvent être identifiées d'éventuelles variations dans le traitement métrique, entre auteurs ou chez un même auteur.

Molière contrairement à Corneille, ne semble pas s'être soucié de réviser ses pièces. Ainsi, trouve-t-on dans les deux premières pièces en vers *L'étourdi* (1653) et le *Dépit amoureux* (1656) des cas du type :

Comme vous voudriez bien, manier ses ducats ;

L'Étourdi ou Les Contre-temps, acte I, scène II.

Et vous devriez mourir d'une telle infamie.

Dépit amoureux, acte V, scène VII.

Où *voudriez* et *devriez* comptent pour deux syllabes. C'est au cours du XVIIème que le décompte des séquences consonne+liquide+I+V se transforme : le *i* acquiert une valeur vocalique et un décompte avec diérèse devient la règle (*de-vri-ez).* Après ces deux pièces, le nouveau décompte sera adopté par Molière.

Le traitement de *oui* et *hier* nous paraît particulièrement intéressant. Ces deux mots partagent une double particularité : ils peuvent constituer une ou deux syllabes métriques ; entraîner ou non l'élision du *e* muet qui les précède. Ils pourraient favoriser une forme d'élasticité du vers. Corneille et Racine ne jouent pas de cette élasticité potentielle : ni l'un ni l'autre ne varient dans le traitement de ces mots.

Commençons par *oui*. Chez les trois auteurs, *oui* est traité comme un monosyllabe. En cela, ils optent pour une prononciation moderne de *oui*. En effet, en ancien français, le *oui* était dissyllabique. D'après Elwert (1965), le *ou* devant voyelle acquiert une valeur consonantique à partir du milieu du XVII^{e} siècle. Cependant le *ou* n'est pas traité de la même manière selon les auteurs. Alors que pour Corneille et Racine, un *e* muet s'élide toujours devant *oui* (le *ou* a donc une valeur vocalique), il est tantôt maintenu, tantôt élidé chez Molière.

Ainsi trouve-t-on dans la même pièce

Notre sœur est folle, oui. / Cela croît tous les jours. (*e* de *folle* élidé devant *oui*, comme chez Corneille et Racine)

Les Femmes savantes, II, IV.

Moi, ma mère ? / Oui, vous. Faites la sotte un peu. (*e* final de *mère* maintenu)

Les Femmes savantes, III, IV.

Sur 38 vers comprenant *oui* après un *e* muet, le *e* muet est élidé 29 fois et maintenu 9 fois. Les cas où le *e* muet est maintenu se situent plutôt dans les comédies les plus burlesques. Le traitement métrique de *oui* est donc instable chez Molière.

Le décompte de *hier* distingue Racine, qui le traite comme un mot dissyllabique de Corneille et Molière qui considèrent le mot comme monosyllabique. Pour Corneille et Racine, le *h* de hier est non aspiré, tout comme chez Molière. Cependant à deux reprises, Molière hésite (ces vers ont été vérifiés dans l'édition d'origine et dans celle du XIX^{e} siècle) :

Et c'est l'homme qu'hier vous vîtes du balcon. (diérèse sur *hier*)

L'École des femmes, II, V.

Et non comme témoin de ce que hier vous vîtes. (*h* de *hier* aspiré)

Dépit amoureux, II, VI.

Enfin, le mot *biais*, absent chez Corneille, traité avec diérèse chez Racine, voit son traitement varier chez Molière : six cas où *biais* est dissyllabique, deux cas où il est monosyllabique.

Et vous deviez chercher quelque biais plus doux. (*biais* dissyllabique)

Molière, Le Tartuffe ou L'Imposteur, V, I.

Voyons, voyons un peu par quel biais, de quel air, (*biais* monosyllabique)

Molière, Le Misanthrope, IV, III

À travers l'étude des variations dans la manière de traiter la syllabe métrique, il apparaît que chez Molière, le traitement du vers est plus relâché et moins cohérent que chez Corneille et

Racine. Tandis que ces derniers sont cohérents de bout en bout sans exception, Molière hésite parfois et fait varier le traitement de trois mots : *oui*, *hier*, *biais*. Seule une exploration systématique pouvait permettre d'identifier *tous* les lieux où il existe bel et bien une variation de traitement de la syllabe. Il s'agit cependant de phénomènes qui ne concernent qu'un faible, voire très faible nombre de vers. Il est donc nécessaire de déployer une approche plus globale.

4. Marqueurs morpho-syntaxiques et accentuels

En nous limitant aux comédies des trois auteurs, nous avons cherché à voir si la répartition des marquages morphosyntaxiques et accentuels produits par le métromètre permettait de différencier les auteurs. Force est de constater qu'il n'y a pas de délimitation nette entre les comédies selon les auteurs si l'on examine un à un les différents critères. Seule *Les Plaideurs* de Racine se distingue nettement de toutes les autres comédies : beaucoup moins d'hémistiches réguliers, moins de mots-outils en début d'hémistiche. Molière se situe entre Racine et Corneille : les hémistiches « réguliers » avec des accents en position paire (type 010101) ou multiple de trois (001001) y sont plus fréquents que chez Racine mais moins que chez Corneille : on ne peut cependant définir des seuils de ruptures. Certains indicateurs marginaux permettent de distinguer les auteurs. Il en est ainsi de « Et » en début de vers. Plus de 17 % des vers de Molière commencent par « Et », contre 11% en moyenne des pièces de Corneille et 7 % des *Plaideurs*.

La répartition des différents indicateurs ne paraît pas être un critère sûr, puisque tous les indicateurs ne convergent pas dans le même sens, et qu'on ne peut définir des seuils. Dès que l'on descend au niveau des pièces (voir annexe), on observe que certaines pièces de Corneille se trouvent dans la zone de Molière et inversement, et cela quels que soient les critères observés. Des effets de sous-genre expliquent sans doute ces variations.

Comédies	Nbre vers	H1 001001	H1 010101	H1 autre	H2 001001	H2 010101	H2 autre	« Et » début vers	Mots-outils en p1	Mots-outils en P2
Corneille	13296	3084	4806	5406	3867	5192	4237	1444	9209	5266
		23 %	36 %	41 %	29 %	39 %	32 %	11 %	69 %	40 %
Molière	17058	3741	5675	7642	4867	6190	6001	2576	11338	7082
		22%	33 %	45 %	29 %	36 %	35 %	15 %	66 %	42 %
Racine	871	139	293	439	203	316	352	42	404	258
		16 %	34 %	50 %	23 %	36 %	40 %	5 %	46 %	30 %
Total	31225	6964	10774	13487	8937	11698	10590	4062	20951	12606
		22 %	35 %	43 %	29 %	37 %	34 %	13 %	67 %	40 %
		chisq =44 ; P<0,0001			chisq=27 ; P<000,1			chisq=210; P<0,001	chisq=49; P<0,0001	chisq=82; P<0,0001

Clef de lecture : Sur les 13296 vers de Corneille, 3084, soit 23 %, ont un premier hémistiche de la forme 001001

Tableau 2. Répartition de quelques critères linguistiques pour les comédies selon les auteurs

Labbé (2001) a montré que *le Menteur* et *la suite du Menteur* de Corneille présentaient de fortes parentés lexicales avec les comédies de Molière. Cette proximité lexicale, confirmée par d'autres méthodes, se retrouve au niveau métrique sur quelques rares critères comme la part de mots-outils en première position, plus élevée que dans les autres pièces de Corneille : il n'y a donc pas forcément de convergence entre les traits liés au vocabulaire et les traits métriques, ces derniers étant assez divergents selon les pièces. Explorons à présent des appro-

ches qui traitent globalement la question du vers, en s'appuyant sur la séquence des syllabes métriques.

5. Modèles statistiques de la syllabe

Dans cette section, nous proposons et évaluons un modèle statistique du vers classique, vu sous la forme d'une séquence de douze syllabes, chacune étant représentée par la suite de phonèmes construite par le métromètre. Cette modélisation statistique vise à répondre à un certain nombre de questions : les différences repérées entre les différents auteurs concernant le fonctionnement de certaines positions métriques peuvent-elles être prise en compte simultanément pour fonder un modèle statistique du vers d'un auteur ? De tels modèles peuvent-il avoir des vertus prédictives pour identifier un genre ou un auteur ? Peut-on en déduire de nouveaux indices permettant de caractériser un type d'écriture ?

Dans un premier temps, nous présentons sommairement le type de modèles probabilistes que nous avons utilisé, ainsi que leur utilisation pour différentes tâches (classification d'un vers ou d'un ensemble de vers). Nous présentons ensuite le corpus de vers sur lequel nous avons travaillé, ainsi que les résultats de diverses expérimentations conduites avec ce corpus.

5.1. *Modèles de langage*

5.1.1. *Bases*

Les modèles de langage sont des modèles probabilistes permettant de modéliser des *séquences d'événements* prenant des valeurs parmi un inventaire fini. Ces modèles ont été particulièrement utilisés dans le contexte de la reconnaissance automatique de la parole, et il existe à leur sujet une littérature abondante, en particulier (Jelinek, 1997), auxquels nous renvoyons le lecteur. Dans le cadre de cette étude, nous avons utilisé les plus simples de ces modèles, connus sous le nom de modèles n-grammes. Soit $S=s_1...s_n$ une séquence, où les s_i appartiennent tous à un ensemble fini V (le vocabulaire), un tel modèle décompose $Pr(S=s_1...s_n)$ selon :

$$Pr(S=s_1...s_m) = Pr(s_1)Pr(s_2|s_1)Pr(s_3|s_1s_2)\dots Pr(s_n|s_1\dots s_{m-1}) \quad (1)$$

$$\approx Pr(s_1)Pr(s_2|s_1)Pr(s_3|s_1s_2)...Pr(s_m|s_{m-n+1}...s_{m-1}) \quad (2)$$

Le passage de (1) à (2), qui caractérise les modèles n-grammes, correspond à une approximation selon laquelle la probabilité conditionnelle d'émission du symbole s ne dépend que des $n-1$ symboles précédents. Par exemple, dans un modèle d'ordre 2 (bigramme), la probabilité d'un événement dans une séquence dépend uniquement de l'événement précédent. Cette approximation permet de ne faire intervenir dans le modèle qu'un nombre fini (de l'ordre de $|V|^n$) de paramètres de la forme $Pr(s|h)$, où h est une séquence d'au plus $n-1$ symboles. Un modèle bigramme est ainsi paramétré par |V| distributions $Pr(./s)$, chacune de ces distributions ayant |V| paramètres.

Par exemple pour modéliser une suite de 0 et de 1 (le vocabulaire V est constitué de deux élements, 0 et 1, |V| =2) dans un modèle bigramme, il suffit de quatre paramètres ($|V|^2$ =4) : la probabilité d'avoir un 0 après 0 ($Pr(0|0)$), 1 après 0 ($Pr(1|0)$) ; un 0 après 1($Pr(0|1)$) et 1 après 1 ($Pr(1|1)$). Il suffit en fait d'estimer deux paramètres, les deux autres se déduisant du fait de la contrainte de sommation à 1.

Les cas d'application de ces modèles au traitement automatique des langues, impliquant la modélisation de séquences de parties du discours, de lemmes ou de formes graphiques, imposent de travailler avec des inventaires de grande taille (V contenant de quelques centaines à

quelques dizaines de milliers de symboles). Même pour de petites valeurs de n ($n=2$ ou $n=3$), le nombre de paramètres est alors exagérément grand. Le problème d'estimation associé en est rendu d'autant plus difficile que, du fait des répartitions très inégale des occurrences de symboles, la très large majorité des séquences ne sont en fait jamais observées, même dans des corpus de très grande taille. En conséquence, estimer $Pr(s|h)$ au maximum de vraisemblance par $Pr(s|h) = n(hs)/n(h)$, avec $n(x)$ le nombre d'occurrences de l'événement x, conduit à affecter une probabilité nulle à de nombreux événements. De nombreuses techniques de lissage de ces distributions, fondées sur l'interpolation de plusieurs modèles ou des stratégies de repli sont donc, dans la pratique, indispensables (Chen et Goodman, 1996).

5.1.2. *Utilisation des modèles stochastiques*

Supposons que l'on dispose maintenant de plusieurs sous-corpus $C_1...C_k$, pour lesquels des modèles $M_1...M_k$ ont été construits. Deux tâches peuvent être distinguées : la *classification* qui consiste à attribuer un vers isolé tiré au hasard à un modèle ; la *segmentation* qui revient à distinguer dans un flux de vers, des sous-séquences homogènes.

Pour la classification, il suffit de calculer, pour toute séquence S de symboles (une séquence étant un vers), le corpus auquel elle se rattache en évaluant le modèle qui rend cette séquence le plus probable.

Pour la segmentation, qui modélise des séquences de vers, il est également possible de construire des segments homogènes, attribuables à un même modèle M_i. Le modèle de segmentation le plus simple consiste à affecter chaque séquence au meilleur modèle, indépendamment des affectations des vers adjacents ; des modèles plus riches permettent de favoriser les fragments homogènes, en pénalisant les changements de modèles. Par exemple, Le rattachement d'une succession de séquences est alors déterminé par programmation dynamique, en calculant les chemins les plus probables dans des automates stochastiques similaires à celui de la Figure 1, qui capture des dépendances d'ordre 1 entre deux modèles M_1 et M_2. Dans ce modèle, la décision de classement d'un vers dépend du classement du vers précédent. En faisant varier les paramètres de ce modèle, il est possible de favoriser l'homogénéité des segments : en particulier, plus $P(M_1|M_1)$ est grand par rapport à $P(M_2|M_1)$, plus il sera « difficile » de quitter le modèle M_1.

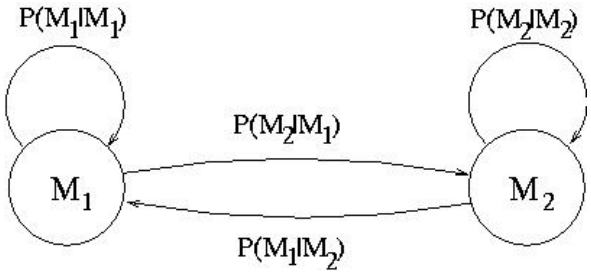


Figure 1. Modèle d'ordre 1 pour la segmentation des pièces

5.2. *Résultats expérimentaux*

Nous décrivons ici les résultats de diverses expérimentations conduites en utilisant ces modèles statistiques : nous décrivons d'abord le corpus utilisé, puis nos expériences de classification de vers isolés, puis de classification et de segmentation de groupes de vers.

5.2.1. *Description du corpus*

Le corpus utilisé contient l'intégralité des alexandrins (du moins ceux reconnus comme tels par le métromètre) des tragédies et des comédies en vers de Corneille, Molière et Racine. Ceci exclut en particulier les vers qui ne sont pas des alexandrins (3 % du corpus) ou encore les vers mal analysés (vers en latin ou en patois). Au total, ce corpus comprend 79456 vers, tirés de 51 pièces différentes[3]. Chaque vers est représenté comme une suite de 12 syllabes. Au total YYY syllabes différentes apparaissent au moins une fois dans ce corpus. Pour les besoins de notre analyse, nous avons extrait intégralement 6 pièces : *Cinna*, *Le Menteur*, *La suite du Menteur*, de Corneille ; *Psyché*, sur laquelle Corneille et Molière ont travaillé ensemble, et *Les Plaideurs*, unique comédie de Racine.

Dans un premier temps, nous décrivons des expériences de classification de vers, qui reposent toutes sur le même protocole, consistant successivement à :

- construire un inventaire des syllabes : dans toutes nos expériences, il s'agit de l'ensemble des syllabes qui ont une fréquence relative supérieure à 10^{-5}. Cet inventaire contient environ 1900 syllabes.
- construire pour chaque sous-corpus un modèle statistique[4] en utilisant 90 % des vers de ce sous-corpus pour l'estimation des paramètres, les 10 % restants étant utilisés pour les tests. La répartition entre apprentissage et test est effectuée en tirant au hasard.
- calculer la probabilité de chacun des vers des corpus de test pour chacun des modèles ainsi construit ; *attribuer le vers au modèle dans lequel il est le plus probable.*

Nous décrivons ensuite des expériences visant à utiliser nos modèles statistiques pour réaliser la classification de vers ou de groupes de vers selon le procédé décrit à la section précédente.

5.2.2. *Classification des vers*

La première question à laquelle nous nous sommes intéressés est celle de savoir si de tels modèles probabilistes étaient à même de capturer des éléments de caractérisation des différents auteurs et genre. À cet effet, nous avons effectué des expériences de classification en sous-divisant le corpus (i) par genre et (ii) par genre et auteur ; puis en utilisant les modèles appris sur ces sous-corpus comme des outils de classification de vers. Les performances de ces systèmes, mesurées en pourcentage de vers bien classés, sont reproduites respectivement dans les Tables 3et 4.

	Comédie	Tragédie
Comédie	1519/60.37	657/13.90
Tragédie	997/39.63	4071/86.10
Total	2516/100.00	4728/100.00

Tableau 3. Classification par genre, avec un modèle 3-gramme : 86 % des vers tirés des tragédies, soit 4071 vers, sont attribués au modèle de tragédie.

[3] *Clitandre*, *Don Sanche d'Aragon*, *Andromède*, *La Toison d'or*, *Tite et Bérénice*, *Pulchérie*, toutes de Corneille, ainsi que la très courte *Pastorale comique* de Molière, qui ne relèvent pas exclusivement d'un seul genre, ne figurent pas dans notre corpus.

[4] Toutes nos expériences ont été menées en utilisant la boîte à outils statistiques SRILM du SRI (Stolcke, 2002). Voir http://www.speech.sri.com/projects/srilm/

Dans le cas de l'identification des genres, on note principalement que si le système classe majoritairement les vers tirés de tragédies comme tragiques, et ceux tirés de comédies comme comiques, les deux genres sont inégalement bien appris : le modèle de tragédie, appris sur près de deux fois plus de vers, fournit un meilleur catégorisateur que le modèle de comédie.

Dans la deuxième expérience, la sous-catégorisation du corpus conduit à construire 4 modèles, estimés sur des ensembles de taille comparable : un pour les tragédies de Racine, un pour celles de Corneille, un pour les comédies de Corneille, et un pour celles de Molière.

	Molière	Corneille (C)	Corneille (T)	Racine
Molière	944/57.35	150/17.24	236/7.94	140/7.97
Corneille (C)	134/8.14	333/38.28	252/8.48	60/3.41
Corneille (T)	410/24.91	297/34.14	1720/57.89	594/33.81
Racine	158/9.60	90/10.34	763/25.68	963/54.81
Total	1645/100.00	869/100.00	2970/100.00	1756/100.00

Tableau 4. Classification par genre et auteur, avec un modèle 2-gramme : 10.3 % des vers tirés des comédies de Corneille, soit 90 vers, sont attribués au modèle de Racine.

Ainsi que ces résultats le montrent sans ambiguïté, les modèles statistiques de type n-gramme capturent des régularités statistiques qui permettent de reconnaître, avec une précision très supérieure au hasard, le genre ou l'auteur de chaque vers de notre corpus. Rappelons que, dans ces expériences, la classification est opérée vers par vers : en utilisant le résultat du classement d'une successions de vers, il devient alors possible d'envisager de rattacher avec très grande précision un groupe de vers à un genre ou à un auteur.

5.2.3. *Classification de pièces*

Les expériences décrites dans cette section visent à répondre à la seconde question, celle de l'attribution d'un groupe de vers à un genre ou un auteur. Pour tenter d'y répondre, nous avons utilisé les pièces initialement mises de côté : pour chacune, nous avons utilisé les modèles d'auteurs construits précédemment pour classer chacun des vers : dans un premier temps les classements sont effectués de manière indépendante ; dans un deuxième temps on pénalise les changements d'auteurs, afin de favoriser les fragments « homogènes », c.-à-d. attribués à un même auteur, en utilisant un modèle d'ordre 2.

Premier exemple : *Cinna*, tragédie Cornélienne par excellence. Pour cette pièce, l'agrégation simple des résultat des classifications individuelles est sans appel : 942 vers, soit plus de la moitié des 1742 vers de la pièce, sont attribués à Corneille [T], Racine se voyant attribuer plus de la moitié (492) des vers restants. Si l'on pénalise les changements d'auteurs, le résultat est encore plus écrasant : 1321 pour Corneille [T], et 322 pour Racine (voir la Figure 2, 0 correspond au premier cas, 1 au second – pénalisation des changements).

L'utilisation de ce même traitement aux autres pièces de notre corpus permet en particulier de constater (i) la position particulière du Menteur (et de *La Suite du menteur*, non représentée ici), qui bien que majoritairement attribué à Corneille, présente un fort pourcentage de vers attribué à Molière ; (ii) la proximité des *Plaideurs* avec les pièces de Molière : on retrouve ici la grande différence métrique entre les tragédies de Racine et cette unique comédie ; (iii) le partage en genre de *L'Illusion comique* (un tiers Comédie, deux tiers tragédie), attribué sans hésitation à Corneille. En ce qui concerne finalement *Psyché*, nous avons pu vérifier que lorsque l'on restreint le modèle de classification à ne choisir qu'entre Corneille et Molière, alors

on retrouve une segmentation conforme à ce qu'on sait de l'écriture de la pièce : les vers attribués à Molière sont très massivement localisés dans le début de la pièce[5].

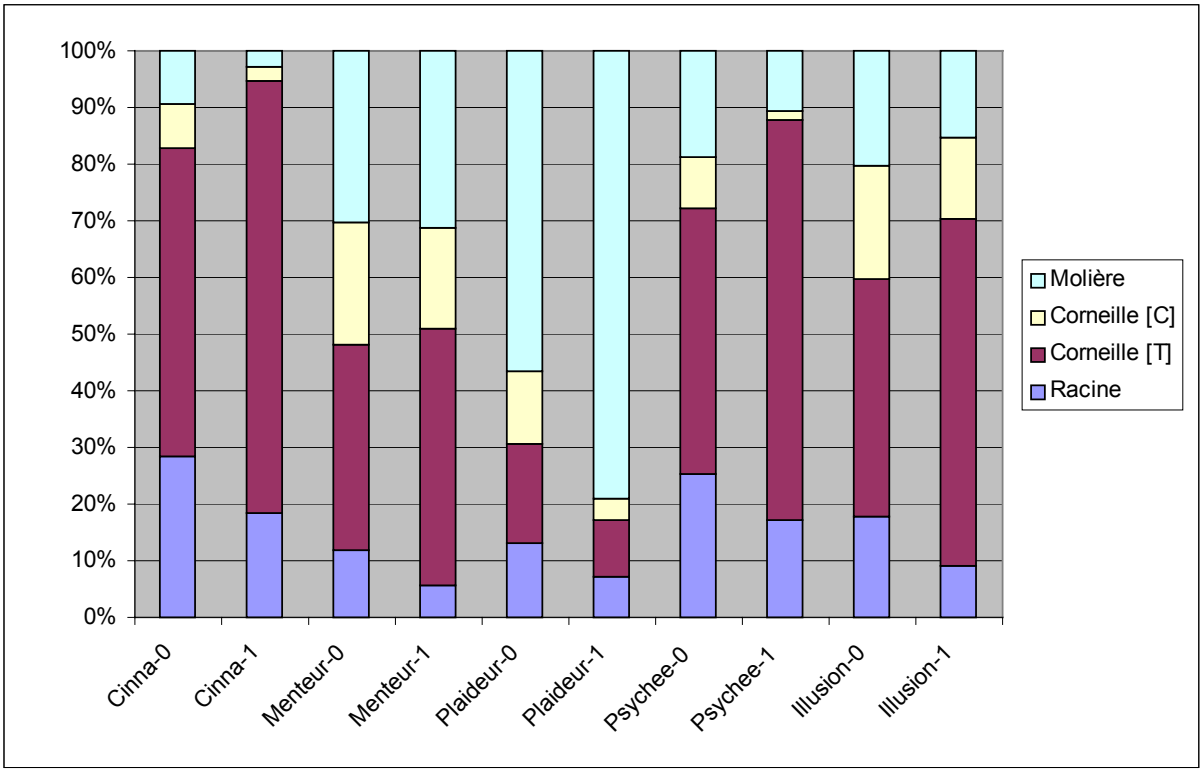


Figure 2. Répartition des vers par modèle pour l'ensemble des pièces

Les résultats obtenus précédemment semblent confirmer l'hypothèse qu'il existe des manières différentes de composer les vers, qui varient suivant les auteurs et les genres (ainsi naturellement à travers les périodes, bien que ce point n'ait pas été abordé ici), et que ces différences peuvent être capturées par des modèles statistiques. Ces résultats ne permettent pas pour autant de conclure définitivement sur la question de l'attribution : rappelons que l'organisation même des corpus d'apprentissage pose comme préalable que ces différences existent ; la modélisation statistique et les expériences ici présentées ne validant ces hypothèses que sous une forme quasi-tautologique : « si ces différences existent, alors elles peuvent être capturées par des modèles statistiques ». Pour répondre plus catégoriquement à ces questions, il faudrait être à même de proposer (et de modéliser) une hypothèse « nulle », consistant à supposer que de telles différences n'existent pas, et à comparer ces deux hypothèses. Ce travail reste encore à faire, mais du moins espérons-nous avoir montré qu'il valait la peine d'être entrepris.

Conclusion

Les aspects métriques constituent des indices dont il est nécessaire de tenir compte dans les travaux de stylométrie. La manière de procéder au décompte des mots peut par exemple être un indice pertinent pour distinguer des auteurs. Pour la première fois, les modèles de langage (modèles probabilistes de séquences d'événements) sont appliqués pour modéliser la séquence des douze syllabes dans le vers pour différents auteurs et genres. Construits sur les

[5] « Ainsi, il n'y a que le Prologue, le premier acte, la première scène du second, et la première du troisième, dont les vers soient de lui [Molière]. M. Corneille a employé une quinzaine au reste », Le libraire au lecteur.

pièces les plus typiques de genres et d'auteurs, ils s'avèrent fournir des modèles efficaces pour la tâche d'affectation : les résultats trouvés semblent montrer qu'il existe bel et bien des modèles de vers différents selon les genres et les auteurs.

Références

Baayen H., van Halteren H., Neijt A. et Tweedie F. (2002). An experiment in authorship attribution. In *Actes des JADT 2002* : 335-346.

Beaudouin V. et Yvon F. (1996). The Metrometer: a Tool for Analysing French Verse. *Literary and Linguistic Computing*, vol. (11/1) : 23-32.

Beaudouin V. (2002). *Mètre et rythmes du vers classique - Corneille et Racine*. Champion, coll. Lettres numériques.

Bernet Ch. (1983). *Le vocabulaire des tragédies de Racine*. Slatkine-Champion.

Chen S. et Goodman J. (1996). An Empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, NM : 310-318.

Constant P. (1991). *Analyse syntaxique par couche*, Doctorat ENST.

Elwert W.T. (1965). *Traité de versification française. Des origines à nos jours*. Klincksieck.

Holmes D.I. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguisitic Computing*, vol. (13/3) : 111-117.

Jelinek F. (1997). *Statistical Methods for speech recognition*. The MIT Press, CA.

Labbé C. et Labbé D. (2001). Inter-Textual Distance and Authorship Attribution Corneille and Molière. *Journal of Quantitative Linguistics*, vol. (8/3) : 213-231.

Muller Ch. (1967, 1992). *Étude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille.* Larousse, 1967, réimpression aux éditions Slatkine, 1979, 1992.

Stolcke A. (2002). SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Speech and Language Processin*g : 901-904.

Yvon Fr. (1996). *Prononcer par analogie : motivation, formalisation et évaluation*, Thèse de l'ENST.

Annexe

		Nb d'alex-andrins	Mots-outils en P1	Mots-outils en P7	« Et » début vers	h1 001001	h1 010101	h1 autre	h2 001001	h2 010101	h2 autre
R	Les Plaideurs	645	63 %	64 %	7 %	17 %	35 %	48 %	25 %	37 %	38 %
C	Mélite	1735	67 %	73 %	9 %	24 %	36 %	40 %	32 %	37 %	31 %
C	La Galerie du palais	1708	67 %	72 %	9 %	20 %	39 %	42 %	30 %	39 %	32 %
C	Le Menteur	1676	72 %	71 %	10 %	22 %	34 %	44 %	26 %	40 %	34 %
M	L'Étourdi ou Les Contre-temps	1875	68 %	73 %	10 %	22 %	34 %	44 %	28 %	38 %	34 %
C	La veuve	1827	67 %	74 %	11 %	23 %	37 %	40 %	29 %	39 %	32 %
C	La Suivante	1602	69 %	74 %	11 %	25 %	36 %	39 %	30 %	40 %	31 %
C	L'Illusion comique	1600	69 %	75 %	12 %	25 %	35 %	41 %	29 %	40 %	31 %
C	La Suite du menteur	1714	73 %	75 %	12 %	23 %	36 %	41 %	27 %	40 %	33 %
C	La Place royale	1434	71 %	75 %	13 %	25 %	37 %	38 %	31 %	37 %	32 %
M	Sganarelle ou Le Cocu imaginaire	599	70 %	73 %	13 %	22 %	36 %	42 %	34 %	35 %	32 %
M	Dépit amoureux	1583	71 %	72 %	14 %	22 %	32 %	46 %	30 %	36 %	33 %
M	Les Fâcheux	760	69 %	70 %	15 %	24 %	33 %	43 %	28 %	36 %	35 %
M	L'École des maris	986	73 %	72 %	16 %	24 %	31 %	45 %	31 %	37 %	32 %
M	L'École des femmes	1584	72 %	71 %	17 %	22 %	32 %	46 %	28 %	36 %	36 %
M	Dom Garcie de Navarre ou Le Prince jaloux	1813	70 %	73 %	17 %	25 %	35 %	41 %	33 %	36 %	31 %
M	Amphitryon	954	72 %	72 %	18 %	22 %	32 %	46 %	30 %	37 %	33 %
M	Les Femmes savantes	1611	75 %	73 %	18 %	22 %	35 %	43 %	29 %	37 %	34 %
M	Le Misanthrope	1676	73 %	73 %	18 %	21 %	34 %	45 %	27 %	37 %	36 %
M	Mélicerte	556	72 %	74 %	19 %	21 %	33 %	46 %	32 %	36 %	32 %
M	Le Tartuffe ou L'Imposteur	1808	75 %	75 %	20 %	21 %	36 %	43 %	27 %	38 %	35 %

Tableau 5. Répartition de quelques critères linguistiques pour les comédies

Analysis of multilingual free responses

Mónica Bécue[1], Jérôme Pagès[2], Campo-Elias Pardo[3]

[1]EIO. Universitat Politècnica de Catalunya – 08028 Barcelona – Spain.
monica.becue@upc.es

[2]ENSA/INFSA – 65 rue de Saint-Brieuc, CS 84215– 35042 Rennes cedex – France
jerome.pages@agrorennes.educagri.fr

[3]Departamento de Estadística. Universidad Nacional de Colombia – Bogotá – Colombia
cepardot@unal.edu.co

Abstract

The analysis of international survey data leads to deal with cross language free responses. To conserve a CA-like approach, which is a reference methodology to analyse open-ended questions, we propose to apply multiple factor analysis for contingency tables. This methodology allows for representing the whole of the lexical tables in a same reference space, keeping CA-like features, in particular transition formulae and interpretation rules. To illustrate the interest of this approach, we use an example extracted from a large international survey in four countries. [1]

Keywords: textual similarity, texts visualisation, cross language open-ended questions.

1. Introduction

The analysis of international survey data with close and open-ended questions leads to deal with free responses expressed in different languages by different samples. A reference methodology to analyse open-ended questions is correspondence analysis (CA) (Lebart *et al*., 1998), applied to lexical tables, aggregated or not (categories_of_individuals × words or individual × words tables). In this work, we want to extend this kind of methodology to several languages in order to tackle multiple aggregated lexical tables.

In textual statistics field, previous works by Lebart (Akuto and Lebart, 1992; Lebart, 1995 and 1998) have proposed a principal axes method to tackle free responses in different languages, allowing for superimposed representations of the categories, as induced by every sample, in a same referential space.

In information retrieval field, various methods have been developed to deal with multilingual corpus. We can cite CL-LSI (Littman *et al.*, 1998) and kernel correlation analysis (KCAA) (Vinokourov *et al*., 2003) which combines LSI (Latent Semantic Analysis which operates a SVD applied to the rough document × term matrix) or ICA (Independent Component Analysis) and canonical correlation analysis for training multilingual retrieval information tools on aligned corpus. These methods allow for representing and comparing the configurations of the words, as induced in every corpus, and searching for equivalent words, in terms of translation, through local similarities among the configurations.

[1] This work has been partially supported by the European project NEMIS (IST- 2001 – 37574/ Information Society Technology Program) and by a grant from National University of Bogotá.

We propose here to use multiple factor analysis for contingency tables (MFACT) which combines features from principal axes methods, canonical analyses and Procustes analysis (Bécue and Pagès, 2004), allowing us to deal with the whole of the responses without translation, relegating this operation to a further step. As in CA, this method offers a visualisation of the distances between categories and also a visualisation of the distances between the whole of the words, meaningful even when the words belong to different languages, being both representations linked by transition rules. Concerning the categories representation, a global representation, as induced by the whole of the answers, and partial representations, as induced by every sample, are obtained, being these representations situated in a same referential space.

2. Data

The data are extracted from a large international survey (Hayashi *et al*., 1992)[2]. People from four countries (Great Britain, France, Italy, Japan) are asked several closed questions and, moreover, the following three open-ended questions:

1 - " What is the most important thing to you in life?"

2 - " Anything else ?"

3 - " What does the culture of your country mean to you?"

The Japanese answers are romanised. We analyse the answers to the first two questions, gathered and considered as a unique response.

In each country, the free answers are grouped into 18 category-documents by crossing gender (male, female), age (into three categories: 18-34, 35-44, 55 and over) and education level (into three categories: low, medium and high). Then, for each country, from the count of words in the whole answers, the lexical table arises by crossing the 18 documents and the most frequent words. Only the words used at least 20 times are kept, this low threshold having been chosen in respect to the 18 categories of respondents.

3. Objectives

The general objective is to compare the structures induced on the categories-rows by the word-columns in the different countries and to detect which categories are similar whatever the countries or not.

Although the word-columns are different for each sample, many of these are the translation of a same word in the different languages. So, we have a special interest in studying their mutual position.

4. Methodology

4.1. *Notation*

The answers given in language t are codified as a (I, J_t) table Y_t in which the element at row i, column j is the relative frequency of word j in the category-document i, as calculated in respect to the sum of terms of the whole of the tables. Word-columns are not the same through the tables. We have T tables (T=4). These tables are juxtaposed row-wise, into a multiple lexical table Y. We denote f_{ijt} the relative frequency, in table Y_t (=country t; t =1,...T), with which row i (=document i ; i=1,...I) is associated with column j (=word j ; j=1,...J_t;

[2] We thank Profesor Lebart to put at our disposal this data set.

$\sum_t J_t = J$). Thus, $\sum_{ijt} f_{ijt} = 1$. We note $f_{i..} = \sum_{jt} f_{ijt}$ the row margin of the table Y, $f_{.jt} = \sum_i f_{ijt}$ the column margin of the table Y, $f_{i.t} = \sum_j f_{ijt}$ the row margin of the subtable Y_t, $f_{..t} = \sum_{ij} f_{ijt}$ the sum of the terms of table Y_t. In the example, T=4; I=18; J_1=106, J_2=96, J_3=55, J_4=48.

4.2. *Analysis of the multiple lexical table*

4.2.1. *Separate analyses*

First, CA is applied to each table Y_t in order to have a first information about common features in the four structures (Figure 1). Then, for every t, t=1,..,T, separate modified margins CA are performed, imposing the row margin $\{f_{i..}, i=1,...,I\}$ and the column margin $\{f_{.jt}, j=1,...J\}$. The first eigenvalue of these pseudo-separate CA, denoted by $\lambda_1{}^t$, will be used in the global analysis to balance the influence of the groups. Performing this CA, with modified margins, is equivalent to perform a general principal axes method on the table with general term given by:

$$\frac{f_{ijt} - \left(\frac{f_{i.t}}{f_{..t}}\right) \cdot f_{.jt}}{f_{i..} f_{.jt}} \qquad (1)$$

using the weights ($f_{i..}$) for the rows (and metric in the columns space) and the weights ($f_{.jt}$) for the columns (and metric in the rows space). In such a way, the rows keep the same weight through all the analyses.

4.2.2. *Global analyses through MFACT*

MFACT (Bécue and Pagès, 2003) consists in a multiple factor analysis (Escofier and Pagès, 1988-1998) extended to contingency tables: a general principal axes method is performed on the juxtaposition of the T tables, with general term given by (1), giving the weight $f_{i..}$ to row i and the weight $f_{.jt}/ \lambda_1{}^t$ to the column (j,t). This method offers results:

- common to all principal axes methods, mainly a global representation of the rows and columns;
- specific to multiple tables, mainly the superimposed representation of the structures induced on the categories by the words in each country, called partial structures, and the representation, in a same reference space, of all the factors obtained in the separate analyses.

5. Results

5.1. *Brief description of the four corpus*

The whole of the free answers of every country constitutes a corpus, identified in the following using the name of the country. Table 1 summarises the main characteristics of the four corpus.

Corpus	Individuals	Corpus length	Distinct words	Kept corpus length	Kept length (%)	Kept distinct words
U. Kingdom	1043	13912	1357	10658	76.6	106
France	1009	14206	1248	11309	79.6	96
Italy	1048	6154	788	4443	72.2	55
Japan	2265	6962	697	5462	78.5	48

Table 1. Main characteristics of the four corpus (Frequency threshold equal to 20)

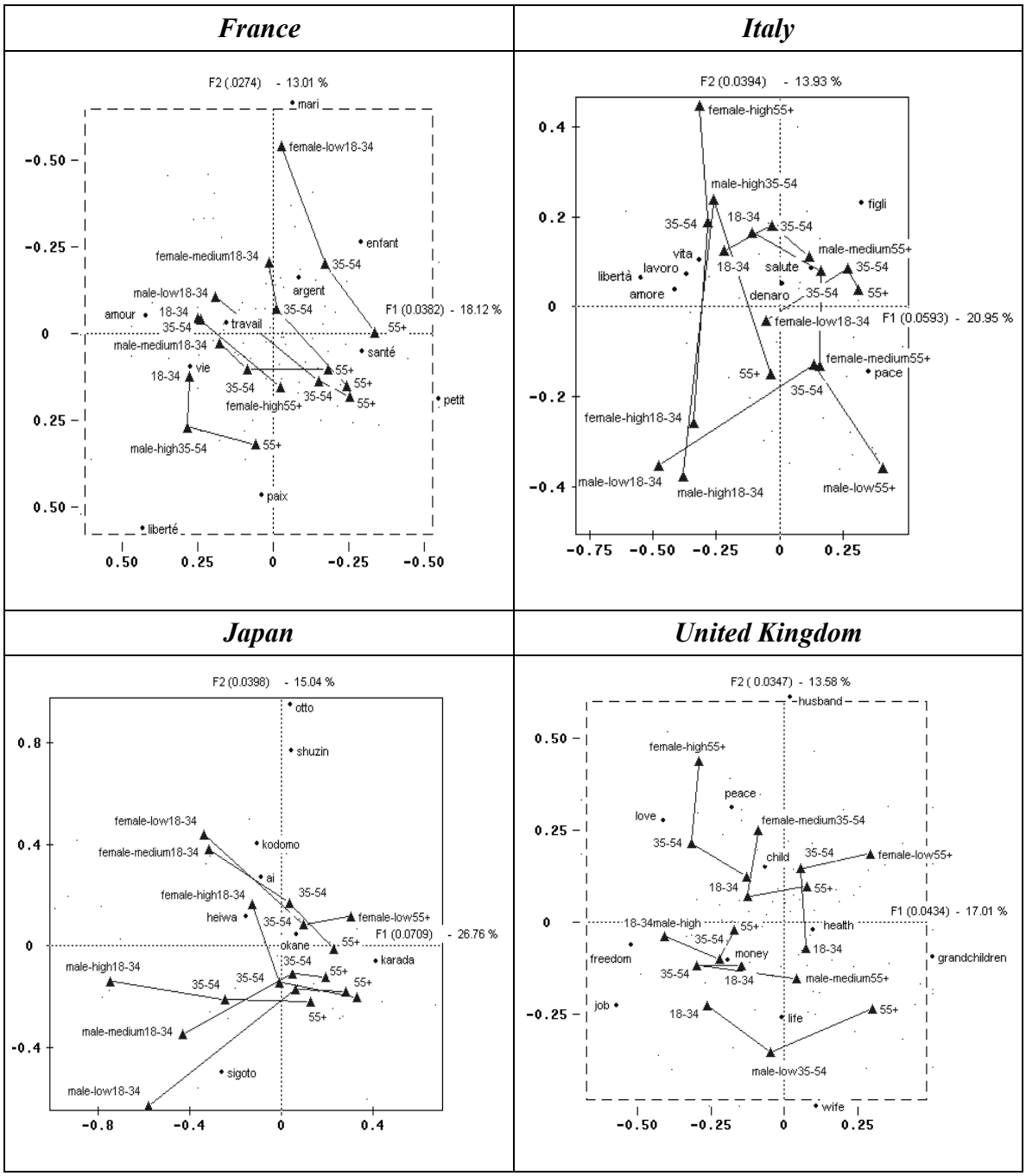


Figure 1. Separate CA in the four countries with trajectories of age, fixing gender and qualification levels

Through the reading of the four lists of words, the frequent topics are identified. Most of them are common to all the countries and are expressed using equivalent words in that sense that they correspond to a same word translated in the four languages.

5.2. *Separate CA*

Figure 1 visualises the four first principal planes, corresponding to the separate CA, in which are drawn the six trajectories of age, fixing gender and education levels.

	Inertia	First eigenvalue	Second eigenvalue	% of inertia kept on the first plane
UK	0.2546	0.0443 (17.4%)	0.0340 (13.3%)	30.7%
FR	0.2149	0.0392 (18.3%)	0.0266 (12.5%)	31.0%
IT	0.2848	0.0570 (20.0%)	0.0415 (14.6%)	34.6%
JA	0.2700	0.0730 (27.0%)	0.0401 (14.9%)	41.9%

Table 2. Inertia and first eigenvalues of the separate CA

The age × gender trajectories present common aspects through all the countries. In the cases of United Kingdom and Japan, age increases along the first axis, while the second axis opposes both genders. In the case of France, there is a structure somewhat similar, but rotated: age increases along the second bissector and the first bissector opposes both genders. Italy looks to be more peculiar, and does not offer, in this first plane, any opposition between genders.

5.3. *Global analysis*

The two first eigenvalues stand out against the following ones: λ_1= 3.52 and λ_2=2.03 (respectively, 18.2% et 10.5% of the total inertia). The high value of the first global eigenvalue (its maximum is equal to 4), indicates that the first global axis is a dispersion direction close to the first separate axis of each subtable, although it is not mixed up.

Table 3-a shows that each corpus contributes, in a balanced way, to the inertia of the first factor. Concerning the second factor, the contribution of Italy is weak, while United Kingdom and France supply the two biggest contributions. The correlations between the first global factor and the projections of the four categories-clouds are strong for all the countries. Concerning the second axis, the correlation is strong with the projections of UK, France and Japan clouds, weaker but nevertheless high in the case of Italy (Table 3.b).

	F1	F2
Total inertia	3.52	2.03
U. Kingdom	0.80	0.76
France	0.93	0.58
Italy	0.88	0.25
Japan	0.90	0.44

Table 3.a. Decomposition per country of the inertia of the two first factors

	F1	F2
U. Kingdom	0.93	0.95
France	0.98	0.93
Italy	0.97	0.81
Japan	0.96	0.90

Table 3.b. Correlations between the projections of the global cloud and the ones of the four partial clouds

It can be concluded that the two first factors are common to the four clouds. The first factor is an important dispersion direction for every country and the second factor is an important dispersion direction for UK, France and Japan, and not so important for Italy.

The visualisation of the categories obtained by MFACT is given by Figure 2. The 6 trajectories of age intervals show a rather regular structure, compromise between the representations offered by the separate CA. Age increases along the first axis, and the second axis opposes both genders, structure similar with that we found in United Kingdom, France and Japan in the separate CA. We can note that the categories with the high education degree have, on the first axis, coordinates which correspond to younger people with lower degrees.

5.3.1. *Interpretation of the proximities between words*

Proximities between words can be interpreted as a resemblance between their users. More precisely, the squared distance between word *j* (belonging to table *t*) and word *k* (belonging to table *r*) can be written in the two following ways:

$$d^2(j \in t, k \in r) = \sum_i \frac{1}{f_{i..}} \left[\left(\frac{f_{ijt}}{f_{.jt}} - \frac{f_{i.t}}{f_{..t}} \right) - \left(\frac{f_{ikr}}{f_{.kr}} - \frac{f_{i.r}}{f_{..r}} \right) \right]^2 \quad (2)$$

$$d^2(j \in t, k \in r) = \sum_i \frac{1}{f_{i..}} \left[\left(\frac{f_{ijt}}{f_{.jt}} - \frac{f_{ikr}}{f_{.kr}} \right) - \left(\frac{f_{i.t}}{f_{..t}} - \frac{f_{i.r}}{f_{..r}} \right) \right]^2 \quad (3)$$

Case 1: the words belong to a same table (t = r)

The proximity between two words is interpreted in term of resemblance between profiles exactly as in the usual CA.

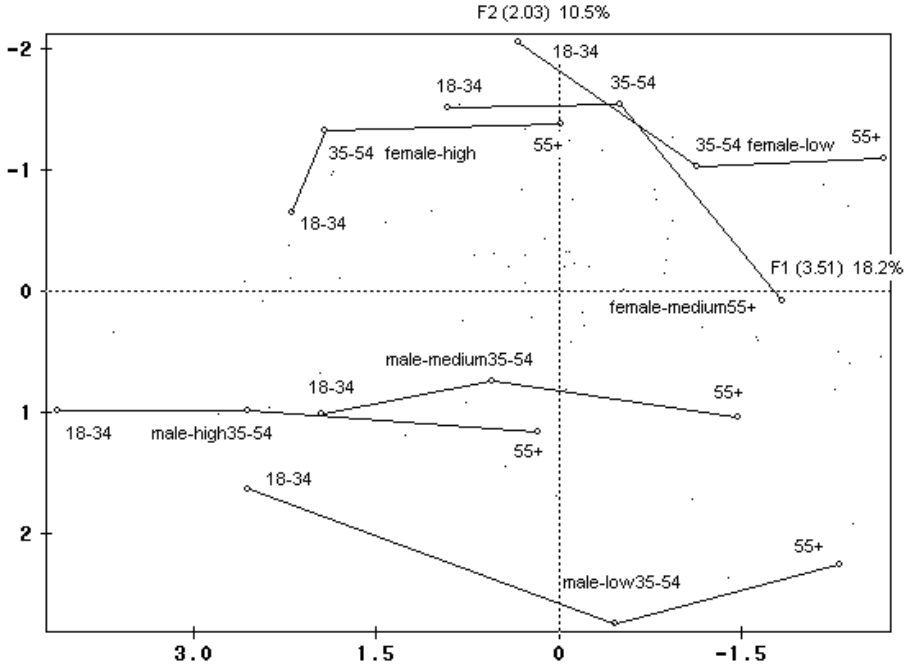


Figure 2 .Global representation of the categories

Case 2: the words belong to different tables (t ≠ r)

The word profiles are relativized by the average profiles, as shown in the two expressions of the squared distance. Expression (2) shows that the profile of a word $\{f_{ijt}/f_{.jt}\ i \in I\}$ intervenes by its deviation from the average profile of the corresponding table $\{f_{i.t}/f_{..t}\ i \in I\}$. Expression (3) shows how the differences between word profiles are relativized by the differences between average profiles, which is important when, as it is the case, the row margins differ from one subtable to another For example, in Italy, the answers of the women between 35 and 54 with a low level of education constitute 4.1% of the Italy corpus, when in United Kingdom corpus the answers of the same category constitute the 10.9% of its length. In fact, the categories have different counts in the four countries.

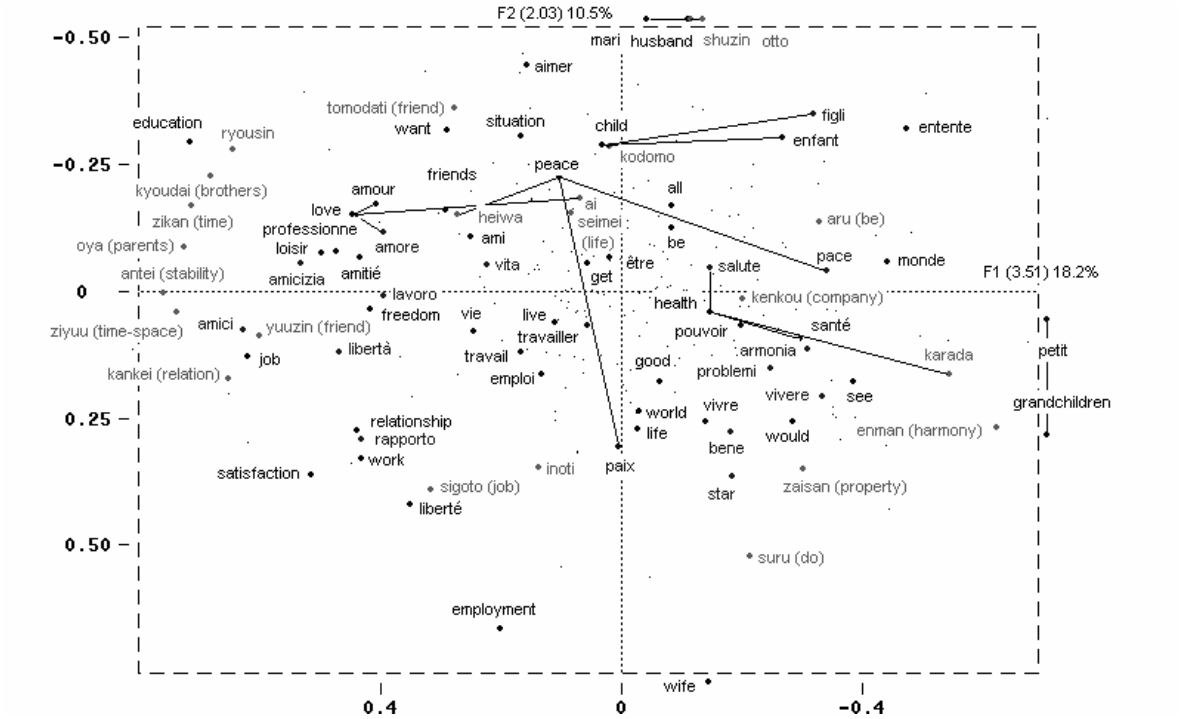


Figure 3. Global representation of the words

Thus, in the example, according to Figure 3, *love*, *amour, amore* are rather quoted by the same categories, when *ai* lies close to the centroid. Consulting the data, we can see that *love, amour, amore* are used, above all by young and medium age individuals (see Table 4) and much more fewer by individuals over 55, while *ai* has a nearly uniform distribution. On the contrary, *peace, paix, pace* and *heiwa* lie in rather distant positions, that corresponds to their different profiles (Table 4).

	Males			Females		
Words	**<35**	**35-54**	**>54**	**<35**	**35-54**	**>54**
love	9 (27.3%)	3 (09.1%)	2 (06.1%)	4 (12.1%)	9 (27.3%)	6 (18.2%)
amour	17 (19.3%)	11 (12.5%)	6 (06.8%)	27 (30.7%)	21 (23.9%)	6 (06.8%)
amore	46 (25.8%)	20 (11.2%)	13 (07.3%)	50 (28.1%)	30 (16.9%)	19 (10.7%)
ai	1 (04.3%)	4 (17.4%)	4 (17.4%)	5 (21.7%)	5 (21.7%)	4 (17.4%)
peace	12 (15.2%)	5 (06.3%)	11 (13.9%)	13 (16.5%)	21 (26.6%)	17 (21.5%)
paix	9 (20.0%)	6 (13.3%)	13 (28.9%)	6 (13.3%)	5 (11.1%)	6 (13.3%)
pace	12 (13.5%)	9 (10.1%)	18 (20.2%)	11 (12.4%)	15 (16.9%)	24 (27.0%)
heiwa	4 (05.3%)	15 (19.7%)	13 (17.1%)	14 (18.4%)	26 (34.2%)	4 (05.3%)

Table 4. Distribution of the words corresponding to love and peace

This fact suggests two kinds of interpretation: the first is linguistic natured (the translation reflects a meaning analogy but not a custom analogy), the second is sociologic natured (the different categories do not have the same values from a country to another).

5.3.2. *Superimposed representation of the partial clouds*

In order to compare the structures induced on the categories by the four clouds of words, we superpose the global description of the categories and those induced by each columns groups.

(partial categories). It allows for pointing out the convergences and divergences between the countries.

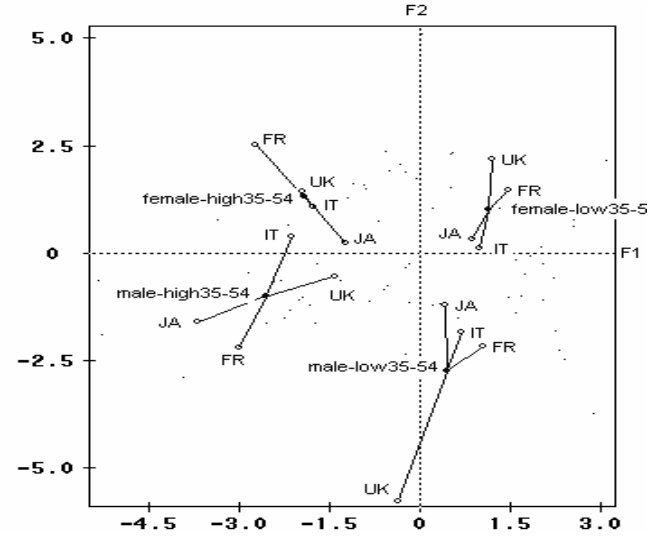


Figure 4. Excerpt of the superimposed representation

For example, the excerpt of the superimposed representation of the partial clouds presented in Figure 4 suggests that that males and females in 35-54 age interval, with high or low qualification, quasi do not differ in Italy. This fact was already pointed out on the separate first planes (Figure 1).

6. Conclusion

The analysis gives a synthetic vision of the diversity of the words used depending on age, gender and qualification. It offers cross language representations of the categories, balanced compromise between the monolingual (or separate) representations, and of the words, being both representations linked by transition rules, which allows us to relate the similarities and dissimilarities between categories to vocabulary and vice-versa. The distances between words from different countries are meaningful and can be interpreted from users profiles. Finally, the superposed representation of the partial categories on the same reference space than the global categories enables to interpret the divergences among homologous categories from a vocabulary point of view.

References

Akuto H. and Lebart L. (1992). Le repas idéal. Analyse de réponses libres en anglais, français, japonais. *Les Cahiers de l'Analyse des Données*, vol. (17/3): 327-352.

Bécue M. and Pagès J. (2004). A principal axes method for comparing multiple contingency tables: MFACT. *Comp. Statistics and Data Analysis* (to be published).

Escofier B. and Pagès J. (1988-1998). *Analyses factorielles simples et multiples; objectifs, méthodes et interprétation*. Dunod.

Hayashi C., Suzuki T. and Sasaki M. (1992). *Data Analysis for Social Comparative research: International Perspective*. North Holland.

Lebart L. (1995). Assessing and comparing patterns in multivariate analysis. In Hayashi C. *et al*. (Eds), *Data Science and application*. Academic Press.

Lebart L. (1998). Text mining in different languages. *Applied Stochastic Models and Data Analysis*, vol. (14): 323-334.

Lebart L., Salem A. and Berry E. (1998). *Exploring Textual Data*. Kluwer.

Littman M.L., Dumais S.T. and Landauer T.K. (1998). Automatic cross-language information retrieval using latent semantic indexing. In Grefenstette G. (Ed.), *Cross language information retrieval.* Kluwer.

Vinokourov A., Shawe-Taylor J. and Cristianini N. (2002). Inferring a Semantic Representation of Text via Cross-Language Correlation Analysis. In *Oral presentation at the 15th NIPS.*

Identification de questions pour traiter les courriels par une méthode question-réponse

Luc Bélanger, Guy Lapalme

RALI – DIRO – Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal – Québec – Canada H3C 3J7
{belanglu, lapalme}@iro.umontreal.ca

Abstract

Automating the process of responding to e-mails is a way of improving customer relation management. We use techniques developed in the area of question-answering for the automatic response to e-mails. This article shows how to identify the questions contained in e-mails coming from the investors relation service of an enterprise. We use a rule-based approach which identifies 81% of the questions found in a corpus of e-mails.

Résumé

La réponse automatisée aux courriels est une solution pour améliorer les services de relations avec la clientèle. Pour y arriver, nous utilisons des techniques issues du domaine de la question-réponse. Nous présentons comment identifier les questions contenues dans un corpus de courriels provenant d'un service de relation avec les investisseurs d'une grande entreprise. Avec une approche à base de règles, nous avons identifié 81% des questions du corpus de courriels.

Keywords: automatic email response, question-answering, question extraction, customer relations management, digital reference.

Mots-clés : réponse automatisée au courriel, question-réponse, extraction de question, gestion des relations avec la clientèle, référence numérique.

1. Introduction

Ces travaux ont été réalisés dans le cadre du projet Mercure dont le but est d'élaborer de nouvelles méthodes de traitement des courriels pour la réalisation de la réponse automatisée aux courriels. Le domaine d'application du projet est le service de relation avec les investisseurs d'une compagnie publique. Les courriels envoyés aux services de relation avec les investisseurs ont la particularité de contenir une grande proportion de courriels qui sont des questions ou qui expriment un besoin d'information s'apparentant à une question. L'approche choisie pour traiter le problème est d'utiliser les techniques issues du problème de la question-réponse. L'identification des questions est une tâche préalable et essentielle pour traiter les courriels par les techniques de la question-réponse car elle permet d'isoler une partie importante du discours nécessaire pour retrouver l'information demandée.

Notre tâche principale a été la création d'une grammaire identifiant les patrons lexicaux les plus susceptibles d'être une question ou une requête. Les patrons ont été conçus manuellement pour obtenir un taux de rappel élevé, au détriment d'un taux de précision plus faible et d'une identification incomplète de l'information pour le traitement automatique. Nous avons pu ainsi repérer 81% des courriels contenant des questions.

2. Présentation du problème

La réponse automatisée aux courriels est nécessaire pour les départements de gestion des relations avec la clientèle des grandes entreprises. Les communications par courriels sont devenues si importantes et si abondantes qu'il est maintenant indispensable d'avoir des moyens pour les traiter automatiquement afin d'améliorer le service. Des études (Jupiter Communications, 2000 ; Banter, 2001) démontrent qu'il est essentiel pour les compagnies d'offrir un service de relation avec la clientèle qui traite les communications électroniques. Pour satisfaire les attentes des clients utilisant ce mode de communication, un courriel doit être traité à l'intérieur d'un délai d'au plus 6 heures, rendant l'automatisation du traitement nécessaire.

Différentes approches ont été proposées pour effectuer le traitement automatisé des courriels. Parmi les plus simples, il y a celles où l'expéditeur du courriel doit lui-même faire la classification de son courriel, pour l'envoyer au bon service, selon des adresses de courriels distinctes. Il y a également les systèmes d'auto-réponse, popularisés par les gestionnaires de listes de discussion (e.g. LISTSERV[1], Mailman[2], Majordomo[3]), qui fonctionnent par activation par des mots-clés. Ces approches sont rudimentaires et elles ne peuvent pas être utilisées en pratique pour traiter efficacement le flot de communications par courriel d'un service avec la clientèle.

Une approche populaire dans les entreprises est l'utilisation de systèmes de gestion des courriels. Ces systèmes demandent une réponse manuelle, contribuant à diminuer le temps de traitement des courriels en automatisant certaines fonctionnalités : la catégorisation des courriels, l'envoi d'accusé de réception, le routage des messages, la suggestion de réponses, l'intégration du système dans l'environnement de travail du préposé, l'archivage des courriels et la production de rapports statistique et historique. Le système Kana Response[4] est un exemple de système de gestion des courriels où des traitements sont effectués automatiquement.

Dans le cadre de notre projet, nous nous intéressons aux approches traitant les courriels de manière autonome. Les systèmes de ce type sont la plupart du temps développés spécifiquement pour un client. Puisque le domaine de ces systèmes est restreint, ceux-ci peuvent plus facilement utiliser des techniques sophistiquées. Les composantes les plus intéressantes utilisées par ce type de système sont : les ontologies du domaine, l'utilisation de patrons de textes, raisonnement à base de cas et méthodes de catégorisation.

Dans cet article, nous abordons le traitement des courriels en considérant qu'ils contiennent des questions ; nous nous inspirons des méthodes des systèmes de question-réponse. Les communications provenant des clients sont souvent une question ou une requête. Au même titre que la question-réponse, la réponse automatisée aux courriels est une tâche apparentée à la recherche d'information. Par contre la requête n'est plus seulement une suite de mots-clés ou une question, mais bien un courriel contenant une ou plusieurs questions.

Watanabe *et al.* (2003) ont tenté de solutionner le problème du traitement automatisé des courriels à l'aide d'un système de question-réponse. Ils ont généré des réponses aux courriels envoyés dans une liste de discussion. Ces courriels ressemblent à une conversation car il y a des courriels de questions et des courriels contenant les réponses. L'étape principale de leur système est l'extraction de phrases significatives dans les courriels à l'aide d'un pointage tenant compte de la présence de noms significatifs, de patrons d'expressions typiques d'une question et d'une pondération liée au nombre d'occurrences de la phrase dans les courriels réponse. Une fois

[1] http//www.lsoft.com/
[2] http//www.list.org/
[3] http//www.greatcircle.com/majordomo/
[4] http//www.kana.com/

les phrases significatives extraites, un calcul de similarité est effectué avec la base de courriels réponse pour déterminer quel courriel retourner comme réponse, similaire à un raisonnement à base de cas.

Les courriels sont des données textuelles avec des propriétés différentes de celles utilisées pas les systèmes de question-réponse élaborés pour les compétitions TREC du NIST (Voorhees, 2001). Les questions utilisées par les systèmes de question-réponse TREC sont extraites à partir des relevés de requêtes envoyées aux engins de recherche, et ensuite nettoyées pour les rendre plus facilement intelligibles pour les systèmes, elles sont généralement courtes et précises. L'information contextuelle contenue dans les courriels est une différence qui doit être exploitée pour le traitement des courriels. Les questions contenues dans les courriels sont énoncées différemment, les formules de politesse et les énoncés d'intention compliquent l'identification et la compréhension des questions.

Dans les systèmes de question-réponse, l'analyse et la classification des questions est un élément essentiel pour chercher efficacement une réponse à la question (Hermjakob, 2001). Les courriels étant des données brutes, nous devons les catégoriser selon l'objet du courriel et la présence de questions. Les données utilisées pour l'élaboration de notre module d'identification des questions sont composées de 210 courriels préalablement nettoyés, analysés et catégorisés manuellement. Ces courriels contiennent tous au moins une question et ils sont rédigés en anglais. La catégorisation des courriels a été réalisée selon le sujet principal de la requête pour déterminer s'il y avait un lien entre le sujet et la syntaxe de la question. Les catégories qui ont été choisies sont : *contact*, *date*, *divers*, *finance*, *invest* et *share*, elles ont été utilisées lors de l'évaluation pour analyser la performance de l'identification. Ces catégories peuvent aussi être utilisées pour identifier la méthodologie à considérer pour répondre aux questions et déterminer quelle information doit être extraite pour trouver une réponse.

3. Solution proposée

L'extraction des questions est réalisée dans un module de traitement de la langue, intégré dans le système d'ingénierie linguistique GATE de l'Université Sheffield (Cunningham *et al.*, 2002), utilisant à la fois de l'information syntaxique et lexicale. Le procédé de détection des questions est exécuté en une série d'étapes. Les premières étapes de traitement sont réalisées par une suite de modules provenant du système GATE, la détection des questions est réalisée par le module d'identification des questions, exécuté en deux étapes. La première étape est l'identification, par un *gazetteer*, des mots dans les courriels qui sont utilisés comme symboles terminaux dans la grammaire. La deuxième étape du traitement est l'identification de patrons de questions, encodée comme une cascade de transducteurs et exprimée dans le formalisme JAPE. Les deux composantes du module sont présentées sous la forme synthétisée d'une grammaire dans la figure 1. Chacune des règles d'identification est numérotée et écrite en *italique*, les symboles écrits en majuscules et en *italiques* sont des symboles non-terminaux et les symboles écrits en police `courrier` sont des terminaux.

Chaque règle a la tâche d'identifier un patron correspondant à une façon de poser une question dans un courriel. La question identifiée correspond à la région du courriel débutant par le début du patron et se terminant à un symbole de fin de phrase, déterminé par le module de segmentation de phrase. Nous décrivons maintenant les règles de la figure 1.

(1) Les patrons `I was wondering` ou `I wonder` de la règle *wonder* sont utilisés dans les courriels pour introduire poliment une question ou une requête formulée indirectement. L'information recherchée pour identifier précisément la requête ne suit pas immédiatement le patron recherché, le patron `if you` *modals action* se retrouve régulièrement entre le

(1) *wonder* → `I` (`wonder` | `was wondering`)
(2) *be_have_mod* → (*MODALS* | *TOBE* | *TOHAVE*)
(`it` | `there` | `these` | *EX*)[*DET*]
(3) *modalsBegin* → (*MODALS* | *TODO*)(*PRP* | *NNP*)
(4) *actionAsking* → *ACTION_ASKING* (`me` | `us`)
(5) *wh_be_have_do* → *WH_WORDS* (*TOBE* | *TOHAVE* | *TODO*)
(6) *Please* → `please` *ACTION_ASKING* (category=VB)
(7) *would_like_to* → [`I` | `we` | `he` | `she` | `and`]`would like to` (category=PRP)
(8) *wh* → *WH_WORDS* [category = PRP]

MODALS → `can` | `could` | `may` | `might` | `must` | `ought to` | `shall` | `should` | `will` | `would`
ACTION_ASKING → `advise` | `forward` | `provide` | `confirm` | `give` | `send` | `direct` | `let` | `tell`
WH_WORDS → `what` | `where` | `when` | `which` | `who` | `why` | `how`

Figure 1. Grammaire d'identification des patrons

patron recherché par la règle et l'information recherchée.

(2) La règle *be_have_mod* identifie des questions portant sur la confirmation d'une information incomplète ou sur la vérification de l'existence de quelque chose (compagnie, méthode de calcul, produit financier, ...) Ces questions sont compliquées à analyser car il faut déterminer l'information cherchée par l'auteur. La création de la réponse est encore plus complexe car elle dépend directement de la confirmation de la requête.

(3) Les questions identifiées par la règle *modalsBegin* sont difficilement catégorisables, le patron utilisé correspond à une syntaxe peu spécifique pour introduire une question. Cette règle pourrait d'une certaine façon être considérée comme un cas général de la règle *wonder*. Les réponses sont déterminées par l'action indiquée par le verbe suivant le patron.

(4) La règle *actionAsking* identifie les questions où le préposé doit exécuter une action. Cette règle sert aussi pour certaines formulations de requêtes similaires aux règles précédentes, mais où la partie d'introduction de la requête est absente ou pas assez fréquente pour être considérée. Les verbes d'actions utilisés pour identifier les requêtes sont : `advise`, `forward`, `provide`, `confirm`, `give`, `send`, `direct`, `let` et `tell`.

(5) La règle *wh_be_have_do* identifie les questions qui débutent par un *wh-words* suivi d'une forme des verbes *be*, *have* ou *do*. Ces questions sont parmi les plus communes, leur forme est généralement celle à laquelle on fait référence lorsque l'on considère le domaine des questions.

(6) La règle *Please* est très similaire à la règle *actionAsking*, la composante importante pour les

deux règles est le verbe énonçant l'action demandée. La différence avec la règle *actionAsking* est que le verbe d'action n'est pas nécessairement utilisé de façon transitive.

(7) Les requêtes identifiées par la règle *would_like_to* sont des questions où le but n'apparaît pas clairement. Le but de la question dépend du verbe suivant `to` dans le patron, celui-ci peut être *ask*, *confirm*, *inquire*, *know*, *attend*, ...

(8) La règle *wh* est réalisée pour récupérer l'ensemble des questions contenant un *WH_WORD* et qui n'ont pas été identifiées auparavant. Les questions identifiées n'ont pas de caractéristiques particulières, elles devront être traitées avec des typologies de questions similaires à ce qui se fait dans systèmes de question-réponse pour pouvoir être répondues.

Lors de l'identification des questions, les règles sont essayées à tour de rôle pour déterminer si le patron de la règle concorde avec le texte. Lorsqu'un patron de règle concorde, l'intervalle de texte débutant au premier mot du patron et se terminant à la fin de la phrase est identifié comme une question. Le mot suivant la fin de la question est ensuite utilisé pour débuter la recherche de la question suivante. L'utilisation de cette stratégie de recherche permet d'appliquer un ordre de priorité sur les règles, qui est l'ordre d'apparition des règles dans la grammaire.

4. Résultats

La méthode d'identification des questions réalise bien la tâche pour laquelle elle a été conçue. En tenant compte du contexte du courriel et du domaine traité, la gestion des relations avec les investisseurs, on constate que les résultats sont très encourageants pour la réalisation de la réponse automatisée aux courriels.

4.1. *Présentation des résultats*

Les premiers résultats sont ceux concernant l'identification des courriels contenant des questions ou des requêtes, ils sont présentés dans le tableau 1, lorsqu'un courriel est considéré comme étant bien identifié, ceci signifie qu'il contient au moins une question identifiée correctement. L'extracteur de questions réalise bien sa tâche, 81% des courriels contenant une ou des questions sont identifiés comme tel. L'identification est particulièrement efficace pour les catégories *date* et *divers*, où les courriels sont identifiés à 95% et 89% respectivement, sans avoir de difficultés avec une catégorie en particulier. L'explication de la bonne performance pour les catégories *date* et *divers* s'explique par le fait que dans la catégorie *date* les questions sont énoncées sous une forme conventionnelle ; dans la catégorie *divers* les questions sont plus faciles à identifier car les courriels sont assurément des questions, où le but est soit multiple, soit carrément hors-catégorie, mais où la syntaxe de la question se conforme aux attentes.

Catégorie	Nbr. courriels	Bien identifiés	Mal identifié
contact	15	11	4
date	24	23	1
divers	19	17	2
finance	61	50	11
invest	32	26	6
share	59	44	15
Total	210	171	39

Tableau 1. Évaluation de l'identification de la (des) question(s) pour chaque catégorie

Le tableau 2 présente la distribution de l'efficacité du module d'identification de questions. Par rapport aux résultats précédents, les données sont maintenant exprimées en fonction de la distribution de la bonne ou mauvaise identification des questions et non plus seulement par rapport à l'identification des courriels. Le taux de succès de l'identification est encore de 81%, même si cette fois on considère le nombre de questions bien identifiées par rapport au nombre total de questions dans le corpus. En examinant les résultats obtenus, on peut constater que les courriels concernant les questions financières de la compagnie ainsi que ceux ayant un lien avec le prix des actions semblent être plus difficiles à traiter. Ceci s'explique par la complexité des énoncés de questions et par l'utilisation du contexte du courriel pour déterminer les sens interrogatifs du message. La catégorie finance possède la plus grande densité de questions annotées par courriel, soit 112 questions réparties à l'intérieur de 61 courriels. C'est la densité du nombre de questions par courriel (≈ 2.35 questions/courriel) qui complique le traitement.

Catégorie	Nombre de questions	Bien identifiées	Mal identifiées	Identifiées à tort			Non identifiées
				normal	sig.	rép.	
contact	16	11	3	1	2	0	2
date	30	27	0	2	3	0	0
divers	49	30	0	2	0	4	12
finance	143	112	10	6	10	2	25
invest	40	41	5	4	1	0	4
share	74	65	4	3	5	0	12
Total	353	286	22	18	21	6	55

Tableau 2. Distribution des identifications par l'extracteur de questions pour les courriels contenant des questions

Les questions identifiées à tort se répartissent en trois catégories dans le traitement automatisé des courriels relativement à l'endroit où elles apparaissent dans le courriel.

1. La première catégorie de questions identifiées à tort est celle où le patron de question apparaît dans le corps principal du courriel, mais où ce patron n'est pas une question. La plupart du temps c'est un *WH_WORD* agissant comme un pronom relatif pour introduire une clause relative, qui n'est pas utilisé dans un contexte interrogatif. L'item (1.1) est un exemple où une question est mal identifiée, l'identification est réalisée par la règle *modalsBegin*.

   (1.1) Please do not hesitate to contact me **should you need clarification or have any inquiries.**

2. La deuxième catégorie de questions identifiées à tort est attribuable aux questions extraites dans la partie signature du courriel, cette catégorie est identifiée par sig. dans le tableau 3. Parfois la phrase extraite doit être considérée comme une question, mais la plupart du temps ce n'est pas le cas, l'identification du patron ne correspond pas à une question où le patron est une formule toute faite du type *Do you Yahoo ! ?*, *Where do you want to go today ?* ou tout autre slogan publicitaire énoncé comme une question.
3. La troisième catégorie de questions qui devrait être ignorée lors de l'identification est celle où la question apparait dans la partie retour (*reply*) d'un courriel (rép.). Ces questions doivent être traitées de façon très particulière, quelques fois elles proviendront d'un courriel redirigé qui doit être répondu, tandis qu'à d'autres occasions ce sera seulement une information qui a suivi le fil d'une *conversation* par courriel.

Le tableau 3 présente pour chacune des catégories de courriel le nombre de questions qui sont annotées par chacune des règles. La distribution des patrons de questions n'est pas uniforme, en fait 270 des 373 questions annotées du corpus le sont avec seulement deux règles (*modalsBegin* et *wh-words*). De plus les règles *wonder* et *actionAsking* ne sont utilisées que pour les catégories de courriels *finance* et *share*, les autres règles semblent être activées de façon uniforme relativement à la quantité de questions des catégories de courriels. Ces résultats mettent en évidence une différence importante entre le traitement automatisé des courriels et les systèmes de questions-réponses de type TREC, les questions ne sont pas toujours introduites par un *WH_WORD* et la syntaxe peut être très diversifiée.

Catégorie	*wh*	*modalsBegin*	*Please*	*actionAsking*	*would_like_to*	*be_have_mod*	*wonder*	*wh_be_have_do*	Total
contact	4	7	2	1 sig.	3	2	0	0	19
date	8	22	5	1 sig.	2	0	0	0	38
divers	9	15	3	0	0	7	0	0	34
finance	87	32	4	13	1	4	0	0	147
invest	15	22	6	0	5	3	6	0	51
share	18	31	15	9	7	2	2	0	84
Total	141	129	35	24	18	18	8	0	373

Tableau 3. Distribution des patrons de questions retrouvées dans chacune des catégories

4.2. *Questions non identifiées*

Selon les résultats présentés auparavant, la grammaire d'extraction réalise bien la tâche pour laquelle elle a été construite, par contre elle ne couvre pas tous les cas de questions qui se retrouvent dans le corpus. Les résultats du tableau 2 indiquent que la grammaire n'a pas été en mesure d'identifier 55 questions, ce qui correspond à 15% du nombre total de questions.

Le nombre de questions non identifiées ne doit pas être considéré comme une contre-performance de la méthode d'identification des questions car il est possible d'en expliquer la provenance. Un examen plus approfondi des données montre qu'il y a cinq courriels dans le corpus qui contiennent près de la moitié des questions non-identifiées (26 questions). Ces cinq courriels ne sont pas représentatifs des données du corpus, ils ont une densité de questions très élevée et ils utilisent le formattage du texte et non la syntaxe pour distinguer les questions. Les autres questions non-identifiées sont dues à deux facteurs, le premier est la faible densité des patrons utilisés pour les questions non identifiées ; le deuxième est la présence de structures grammaticales incorrectes, de coquille et de phrases difficiles à comprendre.

5. Conclusion

L'identification des questions dans les courriels est une étape cruciale pour aborder le problème du traitement automatisé des courriels d'un point de vue similaire à celui adopté pour le problème de la question-réponse. Le processus d'identification des courriels peut être considéré comme une étape de classification pour déterminer les courriels qui pourront être traités automatiquement et un préambule à une catégorisation des questions selon d'autres critères comme le sujet

de la question ou le type attendu de la réponse (prix, date, endroit).

L'approche par patron de surface est celle qui semble être la plus adaptée à notre problème. Il était impossible d'utiliser des méthodes d'apprentissage pour réaliser l'identification des questions car la quantité de données n'est pas assez grande et le corpus de courriel n'est pas assez diversifié pour trouver des facteurs discriminants assez forts. Les résultats obtenus par le procédé d'identification de questions sont très bons si on tient compte des facteurs d'erreurs mentionnés précédemment. Le taux de rappel des questions pourrait être augmenté par l'ajout de règles et la spécialisation des règles existantes, ceci aurait par contre comme effet de diminuer la précision de l'identification.

La réalisation de l'identification est une des étapes d'un procédé pour répondre automatiquement aux courriels envoyés au service de relations avec les investisseurs. Le but du projet étant de traiter les courriels de leur réception jusqu'à la réponse, les questions identifiées devront être analysées pour ensuite être traitées par un système de question-réponse. Le système de question-réponse devra tenir compte du contexte de la question et de l'information extraite lors de l'analyse du courriel, pour faire suivre la question à un module spécialisé qui aura la tâche de trouver l'information pertinente à la rédaction d'une réponse au courriel.

Références

Banter Inc. (2001). Natural Language Engines for Advanced Customer Interaction. [http ://www.realmarket.com/required/banter1.pdf].

Cunningham H., Maynard D., Bontcheva K. et Tablan V. (2002). GATE : A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Hermjakob U. (2001). Parsing and Question Classification for Question Answering. In *Proceedings of the Association for Computational Linguistics 2001 Workshop on Open-Domain Question Answering* : 17-22.

Jupiter Communications (2000). *E-mail Customer Service : Taking Control of Rising Customer Demand*.

Voorhees E.M. (2001). Overview of the TREC 2001 Question Answering Track. In *Proceedings of the Text REtrieval Conference*.

Watanabe Y., Yokomizo K. et Okada Y. (2003). A Question Answer System Using Mail Posted to a Mailing List. In *Proceedings of the PACLING'03* : 335-342.

Analyser les entretiens sociologiques

Jean-Guy Bergeron[1], Dominique Labbé[2]

[1]Ecole des relations industrielles - Université de Montréal – Montréal – Canada
jean-guy.bergeron@umontreal.ca

[2]CERAT-IEP - Université Pierre Mendès-France – Grenoble – France
dominique.labbe@iep.upmf-grenoble.fr

Abstract

How to analyse interviews? We present a collection of 61 interviews about industrial relations in some firms in Quebec. How to transliterate speeches? standardise spellings? mark up the text? tag every word… Then these normalised and tagged texts must be compared with every day speech of the whole population. An experiment is presented with the help of more than 300 interviews, held with French people, the total of which exceeds 2 millions words. It appears that the Quebeckers favour nominal groups and that they use less adverbs and conjunctions than French people whose speeches seem to be rather tense.

Résumé

On examine les problèmes posés par l'analyse des entretiens sociologiques à l'aide d'un corpus d'une soixantaine d'interviews à propos des relations industrielles dans les entreprises du Québec : normalisation orthographique, balisage et lemmatisation des textes. Il faudrait pouvoir comparer les enquêtés avec la population générale. Nous donnons un exemple de cette démarche à l'aide d'un corpus de plus de 300 entretiens réalisés avec des Français (plus de 2 millions de mots). Il apparaît que les Québécois préfèrent le groupe nominal et qu'ils utilisent moins d'adverbes et de conjonctions et que leurs propos sont nettement moins tendus que ceux des Français.

Mots-clés : entretiens sociologiques, France, Québec, lemmatisation, corpus representatif, statistique lexicale.

L'entretien est l'outil privilégié des sociologues. Sans être totalement empirique, l'analyse des transcriptions est encore largement dominée par les méthodes qualitatives. On en trouvera un exemple dans l'ouvrage de Demazière et Dubar (1997) et dans le « symposium » organisé, autour de cet ouvrage, par la revue *Sociologie du Travail* en 1999. Mais l'idée d'une formalisation nécessaire et d'une mise à l'épreuve des outils tend à s'imposer (Jenny, 1997).

Depuis une douzaine d'années, le premier signataire de cette communication utilise, pour l'analyse de ses entretiens, des logiciels d'analyse du contenu comme « Nu-dist ». Par exemple, il a conduit un dépouillement de ce genre sur une soixantaine d'entretiens réalisés, de 1995 à 1998, dans le cadre d'une étude sur la négociation collective raisonnée au Québec (Bergeron et Bourque, 1996, 1998 et 2003). Mais il est évident que, si les méthodes « caqdas » (Computer Assisted Qualitative Data Analysis Software) introduisent de la rigueur, l'opérateur continue à utiliser une grille d'analyse largement *a priori* — avec le risque d'introduire des biais — et qu'il peut aussi passer à côté de choses importantes qu'il n'aura pas pensé à coder… Les méthodes quantitatives de la statistique lexicale, peuvent-elles suppléer certaines de ces carences et permettre aux sociologues de porter un regard plus objectif sur leurs

textes ? C'est ce que nous allons montrer à l'aide de ces 61 entretiens — impliquant au total plus de 70 locuteurs sans compter les enquêteurs. On examinera d'abord les « coûts » — c'est-à-dire, essentiellement, les contraintes qui pèsent sur la saisie des enregistrements — avant de poser le problème des étalons de référence et des méthodes de calcul. Enfin, on évoquera succinctement la richesse des résultats.

1. Saisie, balisage, étiquetage des corpus. Étalons de comparaison

1.1. *Norme de saisie*

Il n'existe pas à l'heure actuelle de véritable « code » de la transcription de l'oral. On a proposé des règles spécifiques visant notamment à restituer le rythme du propos (Benveniste et Jeanjean, 1987). Bien que développée dans un laboratoire important et publiée il y a près de vingt ans, cette norme ne s'est pas imposée, car elle est complexe et éloignée des conventions usuelles. Comment la faire accepter par les opérateurs qui sont chargés de la saisie de ces entretiens mais à qui l'on demandera de continuer à utiliser des conventions profondément différentes pour le reste de leur travail ?

Ne vaut-il pas mieux utiliser la norme standard quitte à accepter une certaine perte d'information ? Ce standard ne peut être que la norme « sténotypique », telle qu'elle est enseignée dans les écoles de secrétariat. Elle consiste, dans son principe, à faire entrer le mieux possible l'oral dans le lit de Procuste de l'écrit avec d'évidentes faiblesses comme pour la ponctuation où les résultats peuvent varier grandement selon l'opérateur. Si celui-ci privilégie la scansion au détriment de la syntaxe : la virgule représente une brève interruption ; une interruption plus longue est transcrite grâce au point si elle est précédée d'une baisse de l'intonation et par un point d'exclamation ou d'interrogation lorsque l'intonation ou la syntaxe le suggèrent. Si l'opérateur privilégie l'analyse syntaxique et sémantique, il ponctuera en fonction des périodes oratoires, en ayant recours aux trois points (souvent envahissants) quand la période ne se termine pas logiquement (du moins à ses yeux).

D'autres inconvénients paraissent inévitables. Signalons l'exemple frappant du participe passé avec l'auxiliaire « avoir ». Dans leurs propos spontanés, ou dans la vie quotidienne, la plus grande partie des locuteurs négligent l'accord (pour les verbes du 2e et du 3e groupes, le féminin doit s'entendre ; de même pour la liaison quand le 's' du pluriel est suivi par un mot commençant par une voyelle). Eh bien ! les secrétaires corrigent systématiquement et sans en avoir même conscience. Dans le cas d'une exploitation secondaire, on ne dispose généralement pas des bandes magnétiques, de telle sorte que sur ce point — crucial pour la réforme de l'orthographe —, ces corpus ne peuvent renseigner sur l'usage « réel » du français...

Enfin, sur ce premier point, une normalisation des graphies est indispensable, notamment pour les sigles, les abréviations, les noms propres, spécialement les patronymes et les toponymes étrangers. Cette tâche est partiellement effectuée par des automates, mais les interventions manuelles sont nécessairement nombreuses et doivent suivre des règles bien précises…

1.2. *Balisage*

Quand on consulte un corpus d'entretiens saisis par plusieurs opérateurs différents, on constate toujours que l'identification des questions et des réponses n'est pas stable. Par exemple, les questions sont parfois en italiques, parfois précédées de « E : » (« enquêteurs »), quelquefois même sans aucun identifiant… On introduit donc des balises qui délimitent les séquences

du texte : en-têtes, remarques, questions et réponses (voir Labbé 2001 et 2002). Grâce à ces balises, l'opérateur pourra isoler le texte des réponses — c'est ce que nous ferons dans la suite de cet exposé — ou celui des questions, s'il s'intéresse au style de la sociologie… etc. Cette opération ne supprime rien, au contraire, elle facilite le traitement de l'information recueillie.

1.3. *Lemmatisation*

La plupart des mots sont susceptibles d'avoir plusieurs graphies : majuscules ou minuscules, élisions, abréviations… Dans certains entretiens, « monsieur » est écrit en toutes lettres, ailleurs M (suivi ou non d'un point), dans d'autres encore : Mr… Les noms de lieux et de personnes sont transcrits phonétiquement (si l'on n'a pas pris la précaution de les transmettre auparavant à l'opératrice). Les chiffres et les dates sont d'une infinie variété, parfois en lettres, parfois en chiffres, avec une virgule comme séparateur de millier, un blanc ou rien du tout. Par exemple, 1990, '90 mais aussi, plus insidieusement : « I99O » (sur le clavier, les I et O majuscules sont plus facilement accessibles que les chiffres…). Plus généralement, des milliers de mots français ont plusieurs graphies (événement et évènement ; puis et peux, etc.) et la réforme facultative de l'orthographe a encore considérablement augmenté leur nombre.

À l'inverse, des mots différents s'écrivent de la même manière. Par exemple, « je suis » (suivre ou être ?), « l'est » (article + nom ou pronom + verbe ?) ou « prise(s) » : substantif féminin (prise de courant), adjectif « pris » (au féminin), verbe « prendre » au participe passé ou… verbe « priser ». Dans tout texte français, ces homographies touchent plus du tiers des mots.

L'énoncé de ces problèmes contient les solutions : normaliser les graphies (un mot, une seule orthographe) et attacher à chaque mot une étiquette qui l'identifie complètement (entrée de dictionnaire et catégorie grammaticale). C'est en s'inspirant de cette idée que, il y a une quinzaine d'années, a été mise au point une chaîne de traitement du français contemporain (Labbé, 1990). La nomenclature des mots, apprise à l'ordinateur, est systématique (par exemple, en français, les substantifs se distinguent par le genre, donc tous les substantifs doivent se voir affecter le masculin ou le féminin), elle est exhaustive (tous les mots doivent y trouver leur place), elle exclut tout double compte, elle ne comporte pas de catégorie ad hoc, ou fourre-tout, etc. Le principe général consiste à regrouper les flexions d'un même mot sous une forme vedette (« lemme ») auquel est associée une catégorie grammaticale. Ainsi, les conjugaisons d'un même verbe sont groupées sous son infinitif ou les pluriels du substantif sous le singulier ou encore les féminins et pluriels de l'adjectif sous le masculin singulier. Par exemple, « être v. » regroupe toutes les formes conjuguées de ce verbe, tandis que « être n. m. » ne se rencontre que sous le singulier et le pluriel.

Remarquons enfin que la lemmatisation doit être réversible — c'est-à-dire qu'on peut retrouver le texte original, sans altération, à partir du fichier des lemmes et qu'elle ne doit pas comporter d'erreur. Telle est la raison pour laquelle les automates — élaborés il y a 15 ans pour la normalisation et la lemmatisation par D. Labbé — résolvent en moyenne 99% des problèmes, laissant à l'opérateur les quelques cas douteux qu'une chaîne entièrement automatique ne pourrait traiter sans erreur…

Les bénéfices de ces opérations sont multiples. Par rapport aux traitements sur les formes graphiques brutes, la normalisation et la lemmatisation redonnent une existence aux verbes (en rassemblant leurs multiples flexions sous une étiquette commune). On peut retrouver certains

mots comme le point cardinal « est », les substantifs « être », « avoir », « avions »… dont les occurrences sont habituellement noyées dans l'océan des formes verbales homographes. Au-delà de ces avantages, la normalisation et la lemmatisation rendent possibles de nombreuses opérations statistiques dont cette étude donnera quelques exemples. En premier lieu, on peut comparer toutes sortes de corpus entre eux.

1.4. *À quoi comparer ?*

Si tous les entretiens ont été saisis, balisés et étiquetés en suivant rigoureusement la même norme, ils deviennent comparables entre eux. On peut opérer des classifications qui feront apparaître les principales sous-populations et mettront en lumière les caractéristiques, de vocabulaire ou de style, propres à chacun de ces groupes (pour un compte rendu de ces opérations sur le corpus « négociation raisonnée », cf. Bergeron et Labbé, 2000). La limite est évidente : on voit ce qui différencie les sous-groupes mais il est plus difficile de révéler ce qui les unit. Comment retrouver les caractéristiques, communes à l'ensemble des individus interrogés, qui les singulariseraient par rapport au reste de la population ? Il faudrait pour cela disposer d'un étalon de référence, une vaste collection d'échantillons représentatifs des pratiques langagières dans la population générale. Un tel étalon existe pour beaucoup de langues. Historiquement, le *British National Corpus* est le premier apparu au début des années 1990 (voir le numéro 8-4 (1993) de *Literary and Linguistic Computing* et Burnard, 1995). Sur l'oral, voir les articles de Crowdy (1993) et Nelson (1997). Les derniers corpus représentatifs parus concernent le tchèque (Kucera, 2002) et l'écossais (Douglas, 2003).

Il n'existe rien de tel pour le français. Certes, la définition d'un étalon de référence pose de multiples problèmes (voir Biber, 1993). Par exemple, a priori, il y a autant de manière de parler français que de territoires francophones. Non seulement, il faudrait avoir des corpus représentatifs du français parlé, en Belgique, au Canada, en France, au Sénégal, en Suisse… mais ces corpus nationaux devraient aussi comporter des sous-ensembles pour ne pas négliger les différences de parler entre Bruxelles, la Wallonie ou entre le Québec, l'Ontario, le Nouveau Brunswick… Pour l'instant, le seul embryon existant a été réalisé par l'Université de Sherbrooke (Centre d'analyse et de traitement informatique du français québécois), et ce corpus n'est que partiellement étiqueté (www.userb.ca/Catifq/bdts).

Comme on le pressent, ces outils seraient très utiles (voir, par exemple, Habert *et al.*, 1997). Nous allons le suggérer en utilisant pour cela un corpus « de référence » qui n'a pas de prétention à la représentativité. Il s'agit d'une vaste base de données rassemblant toutes les transcriptions de l'oral qui nous ont été confiées, depuis une dizaine d'années, aux fins de lemmatisation et de traitement (voir en annexe, une présentation de ce corpus nommé dans la suite de cette communication : « français oral »). Naturellement, on ne peut tirer de telles expériences que des inférences limitées. Il s'agit plutôt de se rendre compte des difficultés que rencontrerait une entreprise comme celle du BNC, sur le français (enregistrement, transcription, correction, étiquetage, indexation et traitement…). Par exemple, nous n'avons pas évoqué, faute de place, les problèmes juridiques posés par l'entreprise (protection de l'anonymat des personnes, de leur vie privée, de la propriété intellectuelle…). Il s'agit aussi d'avoir un aperçu de ce que pourraient apporter les corpus représentatifs pour les grammairiens, les lexicographes, les traducteurs, les enseignants, les gestionnaires de bases de données, etc.

Nous allons en donner deux exemples : la comparaison des catégories grammaticales et l'étude du vocabulaire caractéristique du corpus « négociation raisonnée ».

2. Comparaison des catégories grammaticales

L'étiquetage systématique des textes rend enfin possible l'étude des parties du discours. En effet, la densité des verbes, des noms ou des mots outils varie en fonction des locuteurs et elle est sensible aux thèmes abordés. Le tableau ci-dessous résume les principaux résultats de la comparaison entre le corpus « négociation raisonnée » et le corpus « français oral ». Dans le cas précis, la comparaison soulève une question supplémentaire : les différences proviennent-elles de spécificités propres au français du Québec ? Par exemple, l'excédent des « mots étrangers » peut facilement être rattaché à la situation particulière du pays et à l'emploi, dans la conversation courante, d'un nombre significatif de mots anglais. Ces emprunts ont fait l'objet d'une étude de l'Université de Sherbrooke (www.userb.ca/Catifq/angliweb). Notre corpus en apporte de nombreuses illustrations dont l'évocation dépasserait le cadre de cette communication.

Écarts dans les densités d'emplois des principales catégories grammaticales entre le corpus « Négociation raisonnée » et le corpus de référence « Français oral »

Catégories	Densité des catégories dans le sous corpus	Comparaison avec le français oral
Verbes	**19.8**	**+2.6**
Formes fléchies	*12.7*	*-7.4*
Participes passés	*3.4*	*+37.2*
Participes présents	*0.1*	*+35.8*
Infinitifs	*3.5*	*+19.1*
Noms propres	**0.7**	**-5.1**
Noms communs	**15.3**	**+13.7**
Adjectifs	**3.8**	**+16.2**
Adj. participe passé	*0.6*	*+142.7*
Pronoms	**18.3**	**-7.0**
Pronoms personnels	*9.4*	*-12.3*
Déterminants	**14.2**	**+12.8**
Articles	*10.0*	*+11.1*
Nombres	*1.8*	*+10.7*
Possessifs	*0.7*	*+4.1*
Démonstratifs	*0.5*	*+41.8*
Indéfinis	*1.2*	*+28.4*
Adverbes	**9.2**	**-23.5**
Prépositions	**12.9**	**+22.8**
Conjonctions	**5.6**	**-22.5**
Mots étrangers	**0.1**	**+19.6**

Tous les écarts sont statistiquement significatifs… Comme il y a quelque 70 locuteurs différents dans le corpus de la négociation raisonnée et plus de 300 dans celui du français oral, les excédents et les déficits ne peuvent provenir de caractéristiques individuelles propres à certains enquêtés.

La première surprise provient du net déficit en adverbes et en conjonctions. Par exemple, là où les « Français de France » utilisent 100 adverbes, les enquêtés québécois n'en mobilisent que 76.5 : près d'un quart en moins ! Parmi les principaux adverbes, seule la construction « ne…pas » (ou « ne… plus », « ne… que », etc.) résiste à peu près à cette érosion. Pour le reste, voici la liste des principaux adverbes significativement sous-employés par les enquêtés québécois par rapport au corpus « hexagonal ». Le classement est fait par ordre de « spécificité » décroissante (cf. plus bas) :

> bon, enfin, puis, bien, alors, non, pas, oui, maintenant, même, tout, trop, forcément, si, cher, là-bas, ici, dehors, justement, peu, jamais, mieux, petit, surtout, très, moins, certainement, partout, quelquefois, franchement, apparemment, combien, dessus, déjà, plus, pourtant, grand, pratiquement, là-haut, simplement, pourquoi, automatiquement, parfois, complètement, uniquement, heureusement, demi, effectivement, voire, au-dessus, encore, spécialement, mal, d'abord, normalement, dedans, aujourd'hui, vraiment, vachement, par-là, toujours, a priori, rarement, notamment, largement, malheureusement, pis, bientôt, ailleurs, comment, presque, obligatoirement, hyper, longtemps, systématiquement, hier, plutôt, souvent, directement, tôt, facilement, bas, suffisamment, énormément, actuellement…

En fait, la plupart des adverbes — de temps, lieu ou manière — entretiennent des liens de substitution avec le groupe nominal (voir Arrivé, 1986). Au lieu de « maintenant », on dira : « à l'heure présente », ou « de manière certaine » au lieu de « certainement », etc. Comme on peut s'y attendre, le tableau ci-dessus suggère que le déficit en adverbes se trouve compensé par l'excédent des adjectifs… Cependant, cette substitution n'est pas sans conséquence du point de vue de la communication : les propos ont un aspect plus accompli et moins tendu dont nous donnerons plus bas quelques exemples ;

— le déficit en conjonctions signale une faible coordination des propos (Antoine, 1958-1962). Parmi les principales conjonctions, seuls « comme » et « lorsque » échappent à ce déficit considérable. Outre « que », voici la liste des principales conjonctions très significativement sous-employés : *donc, et, mais, quand, puisque, parce que, sinon, si, car, ou, soit*…

Ces deux déficits majeurs proviennent-ils des enquêtés ? Par exemple, pour la coordination, les interviewés présentent-ils les choses « à plat », sans trop se soucier de les relier entre elles ? Ou bien s'agit-il d'une caractéristique propre au « français du Québec » qui le différencierait fondamentalement du « français de France » ?

L'excédent considérable des participes passés s'explique très probablement par les conditions particulières de l'enquête. La quasi-totalité des entretiens ont eu lieu après la fin de la négociation et la question essentielle était la suivante :

« Pouvez-vous me raconter de façon assez détaillé ce qui s'est passé [à partir du début de la négociation et] jusqu'à maintenant ? »

En revanche, l'excédent en participes présents et en infinitifs est une caractéristique propre à la plupart des enquêtés québécois. Ces formes verbales sont celles qui se rapprochent le plus du groupe nominal. Elles sont à mettre en corrélation avec l'excédent considérable en adjectifs issus du participe passé : négociation (ou méthode, approche) *raisonnée*, moment *donné*, formation *continue*, etc. Par rapport au verbe fléchi, ces formes verbales « dégradées » présentent l'avantage d'effacer, totalement ou partiellement, l'action et l'agent de celle-ci. La tension s'en trouve diminuée. Le propos tend vers l'accompli.

Le déficit important en pronoms personnels peut être rattaché à ce même phénomène de relative « dépersonnalisation » des propos.

Comme on le voit, la simple comparaison des catégories grammaticales a permis de soulever un problème passionnant ! Si l'absence de corpus « nationaux » représentatifs ne permet pas de conclure en faveur de l'hypothèse régionale, celle-ci se trouve plutôt confirmée par la comparaison des vocabulaires.

3. Comparaison des vocabulaires

Pour comparer le vocabulaire des enquêtés avec celui de la population de référence, plusieurs procédés sont envisageables.

Le plus évident consiste à comparer la fréquence d'emploi de chaque mot dans les deux corpus en appliquant les tests statistiques classiques pour la comparaison des fréquences d'un caractère dans deux populations différentes : khi2, loi normale pour les vocables les plus fréquents (fréquence supérieure à 30), loi de Poisson pour les autres, etc. Outre le recours, toujours malaisé, à des tables, on remarquera que ces instruments ne donnent que des approximations. Ce sont des palliatifs inventés à une époque où l'absence d'ordinateur interdisait que l'on puisse envisager le calcul direct.

La loi normale semble fournir un cadre intellectuel logique : on considère les entretiens comme des échantillons extraits aléatoirement d'une urne de Bernouilli constituée par le corpus de référence. Le prélèvement de l'échantillon n'affecte pas le contenu de l'urne : le tirage d'un mot est un événement indépendant de ceux qui l'ont précédé ou qui le suivent… L'espérance mathématique d'un événement — par exemple : probabilité qu'un mot *X* se trouve *n* fois dans l'échantillon — et la déviation standard autour de cette valeur centrale sont aisément calculables. Mais cela suppose que la taille des échantillons prélevés soit très petite par rapport à la dimension de l'urne (afin que le prélèvement n'affecte pas son contenu). La dimension des corpus disponibles rend une telle hypothèse intenable. Par exemple, la taille du « français oral » est de 2,4 millions de mots et celle de l'enquête « négociation raisonnée » de 410 000 mots. De plus, tout texte en langue naturelle comporte une proportion considérable de mots de faible fréquence : aussi grand que soit le corpus de référence, il contiendra toujours une majorité de vocables apparaissant une fois ou très rarement. Le tirage d'un de ces mots « rares » aura une influence évidente sur le contenu de l'urne et sur les épreuves suivantes.

Il est donc nécessaire d'utiliser la loi hypergéométrique — ou « tirage sans remise » (le tirage d'un vocable modifie son espérance mathématique de figurer dans les tirages suivants) — et d'inclure explicitement le corpus sous revue dans l'urne. Ce calcul est inspiré de celui proposé par P. Lafon pour les « spécificités du vocabulaire » (Lafon, 1984 ; Labbé et Labbé, 1994).

Soit :

– le corpus de référence (A) composé de N_a occurrences (taille en « mots ») ;

– le sous-corpus étudié (B avec $B \in A$) composé de N_b occurrences ;

– un vocable i quelconque de fréquence absolue F_{ia} dans A et F_{ib} dans B.

Si les mêmes lois de composition sont en œuvre dans la population générale (A) et dans la sous-population étudiée (B), alors le vocable i aura une fréquence théorique dans B — ou

« espérance mathématique » (E_{ib}) — qui sera fonction de sa fréquence dans A pondérée par le rapport entre la taille de B et celle de A :

$$E_{ib(u)} = F_{ia} * U \quad avec \quad U = \frac{N_b}{N_a}$$

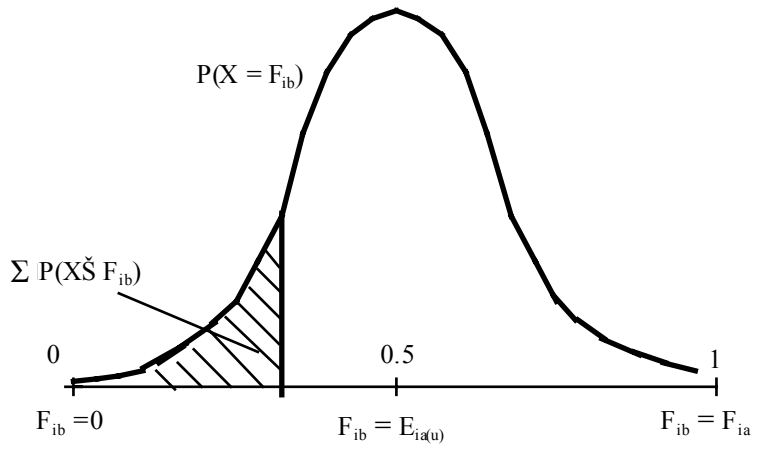
Si la fréquence constatée (F_{ib}) est différente de la fréquence attendue (E_{ib}), quand peut-on dire que le vocable est significativement sur-employé ou sous-employé dans B par rapport à A ? Pour répondre à cette question, il faut considérer la probabilité de l'événement observé F_{ib} par rapport à l'événement attendu (E_{ib}). Cette probabilité suit une loi hypergéométrique de paramètres F_{ia}, F_{ib}, N_a et N_b:

$$(1) \quad P(X = F_{ib}) = \frac{\begin{bmatrix} F_{ia} \\ F_{ib} \end{bmatrix} \begin{bmatrix} N_a - F_{ia} \\ N_b - F_{ib} \end{bmatrix}}{\begin{bmatrix} N_a \\ N_b \end{bmatrix}}$$

F_{ib} peut varier entre 0 — aucune occurrence du vocable dans B — et F_{ia} : toutes les occurrences du vocable sont observées dans le sous-corpus ($0 \leq F_{ib} \leq F_{ia}$).

En développant (1), on constate que le calcul n'a de sens que si $F_{ia} < N_b$ et $F_{ia} < (N_a - N_b)$. La seconde borne va de soi (l'urne doit être nettement plus grande que le corpus sous revue). La première borne signifie que le calcul doit porter sur de grands corpus, ou que, si B est petit, on doit exclure du calcul les vocables les plus fréquents (les « mots-outils »).

À condition que N_a, F_{ia} et N_b soient suffisamment grands, les valeurs de cette probabilité se distribueront selon la fameuse courbe en cloche, avec un mode pour $F_{ib} = E_{ia(u)}$ (graphique ci-dessous).



En arrêtant le cumul des valeurs de P lorsque $X = F_{ib}$, la probabilité pour que le vocable i suive la même loi de fréquence dans le sous-corpus B et dans le corpus de référence A sera la surface comprise sous la courbe (S). Celle-ci varie entre zéro (le vocable n'est pas attesté dans B) et 1 (toutes les occurrences du vocable i se manifestent dans B).

Un vocable sera significativement sur-employé dans B lorsque S aura une valeur supérieure à .975 ou à .995 suivant que l'on choisira un risque d'erreur de 5% ou de 1%. La liaison entre B et le vocable i sera d'autant plus forte que S sera plus proche de 1. A l'inverse, une valeur

inférieure à .005 (ou à .025 si l'on choisit un risque d'erreur de 5%) signifiera que le vocable i est significativement sous-employé dans le corpus sous revue.

Avec cette méthode, le vocabulaire spécifique au corpus « négociation raisonnée » apparaît clairement. Voici par exemple, les substantifs et les adjectifs les plus caractéristiques de ce corpus (avec une chance d'erreur inférieure à 1 sur 1000) :

> *Substantifs* : négociation, problème, gens, chose, temps, partie, syndicat, comité, solution, travail, façon, convention, formation, moment, niveau, personne, intérêt, table, point, employeur, employé, confiance…
>
> *Adjectifs* : raisonné, syndical, bon, patronal, collectif, traditionnel, donné, nouveau, capable, autre, différent, important, difficile, clair, général, long, facile, partiel, intéressant, prêt, évident, conjoint…

La négociation raisonnée (entre les parties syndicales et patronales) est donc le premier thème ; le second tourne autour des « problèmes » (avoir…, poser…, régler…).

En effet, la même méthode peut être appliquée aux syntagmes répétés (Pibarot et Labbé, 1998). Elle fait apparaître le troisième thème principal de ces entretiens, thème qui tourne autour de la « confiance ». Il s'agit notamment de la confiance dans le fait que l'autre partie respectera les règles du jeu, mais aussi la confiance des mandataires — l'employeur d'un côté, les membres du syndicat de l'autre — envers les représentants et, surtout, la confiance en soi-même, dans la justesse et la solidité de ses positions…

Cette même étude appliquée aux groupes verbaux fait apparaître un déficit très net en « modalisateurs » — les verbes pseudo-auxiliaires suivis d'un infinitif (sur le modèle « pouvoir faire, aller voir… ») — sauf, justement, pour « aller » et « pouvoir ». Les principales spécificités négatives sont : « savoir », « croire », « falloir », « vouloir », « devoir ». L'abondance relative en constructions « verbe + verbe » est généralement la marque d'un discours tendu. Autrement dit, sauf pour les modalités orientées vers l'action et le possible, les Québécois semblent nettement moins « tendus » que les Français…

Cette conclusion peut sembler paradoxale. En Amérique du Nord, la négociation collective passe pour très conflictuelle. En effet, le droit du travail n'a pas du tout l'extension qu'il a en France : l'essentiel des droits des salariés sont définis par les contrats collectifs. Les enjeux de la négociation sont donc importants, on discute souvent très longtemps, de manière animée ; la plupart des grèves surviennent à cette occasion. Alors ? la faible tension du discours serait-elle le résultat des méthodes raisonnées qui seraient parvenues à « détendre l'atmosphère » dans des rencontres pourtant traditionnellement conflictuelles ? Ou plus probablement, serait-ce la traduction de traits culturels profondément différents des deux côtés de l'Atlantique ? Seule la constitution de corpus représentatifs permettra de résoudre cette intéressante énigme ainsi que beaucoup d'autres, beaucoup plus fondamentales, concernant notre langue.

Reférences

Antoine G. (1958 et 1962). *La coordination en français*. D'Atrey.

Arrivé M., Gadet F. et Galmiche M. (1986). *La grammaire d'aujourd'hui. Guide alphabétique de linguistique française*. Flammarion.

Berger G. et Leselbaum N. (dir.) (2002). *La prévention des toxicomanies en milieu scolaire : éléments pour une évaluation*. CNDP.

Bergeron J.-G. et Bourque R. (1996). L'impact de la formation sur les pratiques de la négociation raisonnée. In Bélanger J. *et al.*, *Innover pour gérer les conflits*. Presses de l'université Laval.

Bergeron J.-G. et Bourque R. (1998). La formation et la pratique de la négociation collective raisonnée au Québec : esquisse d'un bilan. In Deschênes *et al.*, *Négociation en relation du travail. Nouvelles approches*. Presses du l'Université du Québec.

Bergeron J.-G., Bourque R. et White F. (2003). Empirical Assessment of an Interest-based Bargaining Training Program in Labor-Management Relations. À paraître dans *Relations industrielles*.

Bergeron J.-G. et Labbé D. (2000). L'évaluation de la négociation raisonnée par les acteurs. Une analyse lexicométrique (XVI[e] Congrès de l'AISLF, Québec, juillet 2000). Reproduit dans Bernier C. *et al.*, *Formation, relations professionnelles à l'heure de la société-monde*. L'Harmattan - Les Presses de l'Université Laval. 2002 : 239-252.

Blanche-Benveniste Cl. et Jeanjean C. (1987). *Le français parlé. Transcription et édition*. Didier.

Burnard L. (1995). *Users Reference Guide for the British National Corpus*. Oxford University Computing Service.

Crowdy S. (1993). Spoken Corpus Design. *Literary and Linguistic Computing*, vol. (8-4) : 259-266.

Demazière D. et Dubar C. (1997). *Analyser les entretiens biographiques. L'exemple des récits d'insertion*. Nathan.

Douglas F.M. (2003). The Scottish Corpus of Texts and Speech : Problems of Corpus Design. *Literary and Linguistic Computing*, vol. (18/1-2) : 23-37.

Habert B., Fabre C. et Issac F. (1998). *De l'écrit au numérique*. Masson.

Jenny J. (1997). Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine. État des lieux et essai de classification. *Bulletin de Méthodologie Sociologique*, vol.(54) : 64-122.

Kucera K. (2002). The Czech National Corpus : Principles, Design and Results. *Literary and Linguistic Computing*, vol. (17/2) : 245-258.

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Slatkine-Champion.

Labbé D. (1990). *Normes de saisie et de dépouillement des textes politiques*. Cahier du CERAT.

Labbé C. et Labbé D. (1994). *Que mesure la spécificité du vocabulaire ?* CERAT. Repris dans *Lexicometrica*, vol. (3), 2001.

Labbé D. (2001). Normalisation et lemmatisation d'une question ouverte. Les femmes face au changement familial. *Journal de la Société Française de Statistique*, vol. (142/4) : 37-57.

Labbé D. (2002). *Analyse des représentations du confort électrique à partir d'un corpus d'entretiens* (Rapport pour le GREST-EDF). CERAT.

Meurman-Solin A. (2001). Structured text corpora in the study of language variation and change, *Literary and Linguistic Computing*, vol.(16) : 5-27.

Nelson G. (1997). Standardizing Wordforms in a Spoken Corpus. *Literary and Linguistic Computing*, vol. (12/2) : 79-85.

Pibarot A. et Labbé D. (1998). Les syntagmes répétés dans l'analyse des commentaires libres. In *Actes des JADT 1998* : 507-516.

Pionchon S. (2001). *Les Françaises et la politique* Thèse pour le doctorat de science politique. Institut d'Étude Politique.

Annexe

Le corpus « français oral »

(voir en bibliographie les ouvrages correspondants)

Les Français(es) et la politique (Pionchon) :
32 entretiens : 345 752 mots, 6 540 vocables différents

La négociation raisonnée au Québec (Bergeron et Bourque, 1996, 1998 et 2003) :
61 entretiens : 409 225 mots, 6 591 vocables différents

La prévention des toxicomanies en milieu scolaire (Berger et Leselbaum) :
15 entretiens : 92 992 mots, 4 255 vocables différents

Confort électrique réalisé par les sociologues du Grets-EDF en six enquêtes (Labbé, 2002) :
201 entretiens : 1 270 307 mots, 10 904 vocables différents

Questions ouvertes dans un sondage auprès des femmes divorcées réalisé par l'INED (Labbé,2001) :
3000 enquêtés : 56 107 mots, 2 786 vocables différents

Questions ouvertes dans un sondage auprès des citoyens belges sur la droite et la gauche :
1000 enquêtés : 22 294 mots , 1 706 vocables différents

Divers :
6 entretiens : 115 494 mots, 4 922 vocables différents

Total transcriptions de l'oral :
322 entretiens et deux sondages : *2 264 498* mots, 16 809 vocables différents.

Hasards de la rime

Charles Bernet

UMR 5191 ICAR / ENS-LSH – 15, Parvis René Descartes –
B.P. 7000 – 69342 Lyon cedex – France
charles.bernet@ens-lsh.fr

Abstract

This essay brings into play the analysis tools of quantitative linguistics. The corpus includes a collection of seventieth century French dramatic works, and particularly versified comedies by Corneille, Molière and Racine.

Our purpose is the characterization of rhymes with respect to lexical and phonological data. The lexical analysis relates to the stock of types and tokens concerned in the plays and to the calculus of the inter-textual distance between the various comedies. Tests of signifiance with phonological data, that is frequency counts of syllabes, are used to measure differences and similarities between works or groups of works.

Résumé

Cette communication met en œuvre les outils de la linguistique quantitative. L'étude porte sur un ensemble de pièces de théâtre du dix-septième siècle, et spécialement sur des comédies en vers de Corneille, Molière et Racine. Son objet est l'analyse de la rime du point de vue lexical et phonologique. L'analyse lexicale porte sur le nombre de mots-formes et de vocables employé à la rime dans les pièces et sur le calcul de la distance intertextuelle entre les comédies. Les données phonologiques, c'est-à-dire les syllabes placées à la rime, donnent lieu à des tests statistiques pour établir des différences et des similitudes entre les pièces et groupes de pièces.

Mots-clés : linguistique quantitative, stylistique, vers, rimes, théâtre français, dix-septième siècle, Corneille, Molière, Racine

1. Objet de l'étude

Dans les textes en vers, les mots placés à la rime méritent une attention particulière parce qu'ils sont, en tant que tels, mis en relief dans la composition du texte. Il ont été sélectionnés à la fois pour leur sens et pour leurs caractères formels et leur examen pourrait révéler des traits propres à une œuvre, un auteur ou un genre littéraire. Selon l'hypothèse qui sous-tend ce travail, le choix des mots-rimes[1] n'est pas le fruit du hasard. Le titre de cette contribution, qui est une allusion à Baudelaire[2], doit donc être compris par antiphrase.

Un volet de cette étude est consacré à l'analyse des mots placés à la rime et un autre à l'examen des syllabes finales de chaque vers[3]. Le corpus comporte vingt comédies de Corneille, Molière et Racine. Pour certaines opérations, ces vingt pièces ont été replacées dans un ensemble plus vaste et plus varié de soixante-huit pièces couvrant l'ensemble du dix-septième

[1] Nous empruntons ce néologisme à V. Beaudouin (2002 : 92).

[2] « Je vais m'exercer seul à ma fantasque escrime, / Flairant dans tous les coins les hasards de la rime , / Trébuchant sur les mots comme sur les pavés » (Charles Baudelaire, *Les Fleurs du mal*, Tableaux parisiens, « Le soleil »).

[3] Une grande partie des traitements statistiques a été réalisée à l'aide de programmes mis au point par Dominique Labbé, notamment pour le calcul de la distance intertextuelle. Le travail sur les syllabes finales a été possible grâce à Valérie Beaudouin qui nous a communiqué les traitements obtenus avec le métromètre. (Beaudouin et Yvon, 1996 : 23).

siècle afin de mettre en évidence des caractères qui pourraient les opposer aux pièces appartenant à d'autres genres dramatiques[4].

2. Corpus, normes de lemmatisation

Le corpus comporte vingt pièces en alexandrins de Corneille, Molière et Racine, soit les huit comédies de Corneille, onze comédies de Molière et l'unique comédie de Racine. Ont été retenues les pièces qualifiées de comédies par leurs auteurs, à l'exception des trois comédies héroïques de Corneille[5]. *Psyché,* pièce de Molière écrite en collaboration avec Corneille et Quinault, dont l'examen s'imposerait dans une étude spécifiquement consacrée à des questions d'attribution de texte n'a pas été retenue car elle n'a jamais été qualifiée de *comédie* mais tantôt de *tragi-comédie et ballet* et tantôt de *tragédie-ballet*. Enfin, une comédie de Molière*, Amphitryon,* qui n'est pas en alexandrins mais en vers libres, a été écartée.

Les vingt comédies s'échelonnent sur une période qui couvre près d'un demi-siècle. Parmi les comédies de Corneille, un premier groupe de six pièces, antérieur à l'échec du *Cid* s'échelonne de 1629 à 1636. Après une interruption, Corneille revient au théâtre avec quatre grandes pièces tragiques[6] suivies du *Menteur* en 1643 puis de *la Suite du Menteur* en 1644. Son œuvre comique s'arrête là. La date des premières comédies de Molière n'est pas parfaitement établie ; on estime cependant que la première représentation de *l'Étourdi ou les Contretemps* se situe à la fin de l'année 1654, soit une décennie après la *Suite du Menteur*. Les comédies en alexandrins de Molière se succèdent jusqu'en 1666 avec *Mélicerte*, après laquelle les pièces en prose se font plus nombreuses ; sa dernière comédie en vers est *les Femmes savantes* en 1672. La comédie de Racine, *les Plaideurs*, est représentée et imprimée en 1668. Le texte pris en compte ici est celui de l'édition des *Grands écrivains de la France*[7].

Les relevés effectués dans les pièces du corpus enregistrent les derniers mots de tous les vers, quel que soit leur type (alexandrins, décasyllabes, octosyllabes, etc.). Les vers sans rimes ont été écartés ; c'est le cas notamment de 12 vers tronqués dans *Dom Garcie de Navarre* (v. 494-505) et d'un passage en latin, tiré et adapté d'Ovide, dans *les Plaideurs* (« *Unus erat toto naturae vultus in orbe / Quem Graeci dixere chaos, rudis indigestaque moles* », v. 809-810).

En nous appuyant sur la lemmatisation de Ch. Muller (1967) dans son étude sur Corneille, reprise avec quelques modifications dans *l'Étude du vocabulaire des tragédies de Jean Racine* (Bernet, 1983) et dans l'analyse du *Vocabulaire de Molière dans les comédies en alexandrins* (Kylander, 1995), nous avons adopté des principes d'identification du mot compatibles avec les contraintes propres à la versification qui pèsent sur le mot-rime. Un mot ne peut jamais rimer avec lui-même et les mots isomorphes qui peuvent être associés à la rime doivent être considérés comme distincts. Souvent, les isomorphes appartiennent à des catégories grammaticales différentes. Ainsi, le mot *pas,* substantif masculin, peut rimer avec *pas* adverbe de négation ; de même pour *point.* Il arrive aussi qu'ils appartiennent à la même

[4] Les comédies des trois auteurs sont repérables par la typographie dans le tableau 1 figurant en annexe.

[5] Il s'agit de *Don Sanche d'Aragon,* de *Pulchérie* et de *Tite et Bérénice*. La comédie héroïque est définie ainsi dans le *Dictionnaire dramatique* de Laporte et Chamfort (1776) : « On appelle de ce nom une Piéce dont l'intrigue, purement romanesque, est dépourvue de ce comique qui provoque le rire, & dont le dénouement heureux ne coûte ni de sang aux Personnages, ni de larmes aux Spectateurs. Ce genre se soutient par des aventures extraordinaires, des bravades, des sentimens généreux. »

[6] *Horace*, *Cinna*, *Polyeucte* et *Pompée.*

[7] Corneille P. (1862-1868). *Œuvres*, nouv. éd. par M. Ch. Marty-Laveaux. t. 1-12. Hachette. Molière (1873-1900). *Œuvres*, nouv. éd. par M. Eugène Despois. t. 1-14. Hachette. Racine J. (1886). *Les Plaideurs : comédie*, nouv. éd. par M. Paul Mesnard, in *Œuvres*, t. 2. Hachette.

catégorie, c'est le cas du substantif *état* associé à la rime à *État*[8], dans ce cas, ils sont donc considérés comme deux mots différents. Inversement, sont groupées sous un même lemme des formes, appartenant à plusieurs parties du discours qui ne sont pas perçues comme suffisamment différentes pour être associées à la rime (ex. *coupable* ou *malheureux* adjectifs et substantifs ; *net* ou *soudain* adverbes et adjectifs ; *bas* ou *haut* adverbes, adjectifs et substantifs). Notons que les formes appartenant aux catégories du verbe et du substantif (ex. *devoir* ou *souvenir*) conservent deux entrées distinctes. Le résultat est une norme très synthétique[9], avec laquelle le nombre des vocables différents d'un texte est sensiblement inférieur à celui que l'on obtient lorsque l'identification des unités lexicales repose sur les catégories grammaticales.

Quelques interventions sur les graphies ont consisté à harmoniser des divergences dues à des choix éditoriaux différents. Ainsi, par exemple, les formes *conte, conter* et dérivés (ex. *méconte*) des œuvres de Corneille ont été dégroupées sous *conte*, *conter*, et *compte*, *compter*, etc. de façon à adopter une norme lexicale unique et cohérente pour les trois auteurs. Pour la même raison, les pluriels en *-ans* (*complaisans, enfans, puissans*) et *-ens* (*accidens, indulgens, sentimens, talens*) ont été regroupés avec les pluriels en *-ants* et *-ents* ; certaines finales en *-ette* (*inquiette, discrette, secrette*) ont été regroupées avec les mêmes mots en *-ète*. En revanche, l'omission des consonnes finales correspondant à des licences poétiques (*doi* pour *dois*, *sai* pour *sais*) n'a pas donné lieu à des regroupements ; les raisons de ce choix sont explicitées *infra* au § 6.

3. Richesse lexicale

Un premier groupe d'analyses porte sur la variété des mots-rimes. Les résultats figurant sur le tableau 1 placé en annexe sont obtenus par calcul de la « richesse lexicale » pour les mots-formes ainsi que pour les lemmes. Dans les deux cas, les effectifs théoriques sont calculés en ramenant toutes les pièces à une longueur canonique de 1000 mots-rimes – colonne E(V') – puis de 600 mots-rimes, ce qui correspond à la taille de la pièce la plus courte, *Mélicerte* – colonne E(V''). Les pièces sont classées par ordre de richesse à la rime décroissante.

Les comédies occupent la partie supérieure de la liste et sont donc parmi les pièces qui offrent une grande variété lexicale à la rime. Elles se placent au niveau de tragédies du début du 17e siècle (*Scédase, Lucrèce* et *Hector*) dont on a déjà souligné l'exceptionnelle richesse du vocabulaire (Bernet, 1999 : 193), des premières œuvres tragiques de Corneille (*Médée* et *Clitandre*), des tragédies sacrées de Racine (*Esther* et *Athalie*) et enfin, de chefs-d'œuvre de la maturité de Corneille (*Cinna*, *Polyeucte* et *Pompée*). Chez Corneille, la variété des mots-formes à la rime commence par décroître dans les cinq premières comédies, qui forment un ensemble esthétique et thématique cohérent car elles ont pour point commun de constituer « un reflet d'une société » (Gilot et Serroy, 1997 : 81), entendons par là un reflet de la bonne société parisienne, qui retrouvait dans ces comédies une réalité et un langage soigné qu'elle connaissait bien. Elle remonte ensuite avec *L'Illusion Comique,* qualifiée par Corneille lui-même d'« étrange monstre[10] » car cette pièce est une construction gigogne mêlant plusieurs genres, et avec les deux *Menteurs* qui appartiennent à la tradition, en vogue après 1640, de comédies à la manière de celles du Siècle d'Or espagnol.

[8] « J'en admire beaucoup dont on fait peu d'état ; / Leurs fautes, tout au pis, ne sont pas coups d'État » (Corneille, *La Galerie du Palais*, v. 183-184).

[9] D'autres précisions sur cette norme figurent dans Bernet (199 : 189-190).

[10] Dans son *épître à Mademoiselle M.F.D.R.* (*Œuvres*, éd. citée, t. 2 : 430).

L'étalement des valeurs est plus ample pour les comédies de Molière que pour celles de Corneille. B-M Kylander a déjà montré, à propos des caractéristiques quantitatives de l'ensemble du vocabulaire dans les comédies en vers de Molière, que « chaque pièce présente un profil lexical individuel nettement distinct de celui des autres » (Kylander, 1995 : 237). Notre corpus regroupe exclusivement des comédies, mais il existe à l'intérieur de ce genre une ample diversité qui va, chez Molière, de la farce avec *Sganarelle*, à la comédie pastorale avec *Mélicerte.* Aucun de ces deux genres n'est représenté chez Corneille. Les deux pièces les plus marginales ne sont pas de vraies comédies. *Mélicerte,* spectacle destiné à la Cour, appartient à une tradition qui mettait en scène des bergers et bergères dans un cadre bucolique autour d'une intrigue romanesque. On y trouve à la rime un vocabulaire plus sobre que celui de toutes les pièces proprement comiques et l'on ne doit pas s'étonner de trouver cette pièce proche de *Don Sanche d'Aragon* et d'*Andromède*, deux comédies héroïques de Corneille. L'autre pièce de Molière qui se situe à l'écart est *Dom Garcie de Navarre*. On notera que celle-ci, de la même veine que *Don Sanche d'Aragon*, a été qualifiée de comédie héroïque par ses éditeurs à partir de 1734. Ce fut aussi le seul grand échec de Molière auprès du public, qui préférait ses pièces comiques. Selon B-M Kylander, « la structure quantitative générale du vocabulaire de *Dom Garcie* s'écarte fortement de celle des autres comédies de Molière » (Kylander, 1995 : 187) et la valeur extrêmement basse de l'indice pronominal [indicateur de la familiarité du style] « place cette pièce très loin des comédies et même parmi les tragédies de Corneille et de Racine qui […] ont le style le plus noble et soutenu » (Kylander, 1995 : 189). Notre constat corrobore ces remarques.

La diversité des mots-rimes peut être obtenue soit par l'emploi des mêmes mots affectés de nouvelles variations morphologiques, soit par l'apport de vocables nouveaux. Pour savoir ce qu'il en est dans notre corpus, il convient d'examiner les divergences entre les classements obtenus dans le tableau 1 selon que le calcul est effectué avec les effectifs des mots-formes ou avec les effectifs des lemmes. Pour les pièces comportant plus de 1000 mots-rimes, les écarts de rang, même les plus importants, ne peuvent pas être considérés comme significatifs car les valeurs calculées se situent toujours dans l'intervalle de confiance. Notons cependant le cas de *L'Illusion comique* qui a un meilleur classement selon les vocables que selon les mots-formes, donc un vocabulaire plus riche que que l'on pouvait escompter. Parmi les pièces plus courtes, *Sganarelle* et *les Plaideurs* présentent la même caractéristique.

4. Distance intertextuelle

Ces données permettent aussi d'apporter une contribution au débat sur Corneille et Molière qui a animé notre communauté au cours de ces deux dernières années. Le calcul des distances, selon la méthode de C. Labbé et D. Labbé a été appliqué aux 16 comédies qui comportent plus de 1000 mots-rimes. L'indice atteint des valeurs très supérieures à celles qui sont obtenues pour la totalité du vocabulaire. Alors que, dans le tableau 2 en annexe, l'indice varie respectivement de 0,60 à 0,77 pour les données non lemmatisées et de 0,51 à 0,70 pour les données lemmatisées, il se situe seulement entre 0,18 et 0,33 dans le travail de C. Labbé et D. Labbé (2001 : 225)[11].

L'ampleur des différences indique, dans toutes les comédies, une variabilité considérable des mots placés à la rime. Il ne faut donc pas surestimer l'effet des contraintes qui pèsent sur le

[11] .Rappelons en outre que la lemmatisation des mots-rimes diffère de celle qui a été mise en œuvre par ces deux auteurs.

choix de ces mots ; la taille du lexique en jeu est très importante. On peut s'en convaincre en examinant la répartition des mots-rimes dans les pièces du corpus.

Les mots-formes attestés dans chacune des vingt comédies ne sont qu'au nombre de 15. Il s'agit, par ordre de fréquence décroissante, de *vous, moi, âme, pas* (adv.), *cœur, dire, yeux, jour, elle, bien, lui, ici, envie, colère* et *aujourd'hui*. Les chiffres restent faibles pour les tranches de répartition suivantes : 17 mots-formes ne sont attestés que dans 19 pièces, 28 dans 18 pièces, 19 dans 17 pièces, etc. La répartition des mots-rimes n'est en rien comparable à ce que l'on peut observer pour l'ensemble du vocabulaire car il n'existe pas, pour les mots placés à la rime, de vocabulaire minimal qui se trouverait nécessairement dans n'importe quel texte.

Si les valeurs de l'indice sont plus élevées sur nos données que sur celles qui on été exploitées par C. Labbé et D. Labbé, les écarts d'une pièce à l'autre sont souvent similaires. Globalement, la distance entre les pièces est moindre à l'intérieur de l'œuvre de chaque auteur. Parmi les exceptions, on mentionnera, comme le montrent les résultats du tableau 2, la singularité des deux *Menteurs* parmi les pièces de Corneille et celle de *Dom Garcie* parmi celles de Molière.

Le Menteur et la *Suite du Menteur* sont relativement éloignées des autres pièces de Corneille sans être pour autant plus proches de celles de Molière, à l'exclusion de *l'Étourdi* et du *Dépit amoureux.* La singularité des comédies du *Menteur,* au regard de la statistique lexicale, a été soulignée par Ch. Muller qui note à leur propos : « Chaque fois que l'on inscrit dans l'ordre chronologique un indice quelconque, deux chiffres sortent du rang, et il n'est pas besoin de consulter les titres pour déceler une rupture dans le style » (Muller, 1967 : 213). Mais Corneille lui-même en avait déjà averti ses lecteurs ; l'épître qui introduit *le Menteur* commence ainsi : « Je vous présente une pièce de théâtre d'un style si éloigné de ma dernière [*i.e. Pompée*], qu'on aura de la peine à croire qu'elles soient parties toutes deux de la même main, dans le même hiver. » (Corneille, *Œuvres*, éd. citée, t. 4 : 130).

Afin de mettre en évidence les convergences entre les mots-rimes des *Menteurs* et ceux des comédies de Molière, on peut examiner tous les vocables en intersection entre ces deux ensembles et absents des autres pièces de Corneille.

Un premier groupe est lié au mensonge et aux artifices : *mentir*[12] (12 – 2), *menti* (2 – 1), *mens* (1 – 1), *menteurs* (1 – 1), de même que de *déguisements* (2 – 1), *grimace* (2 – 10), *déguisé* (1 – 1), *embuscade* (1 – 1), *imposteurs* (1 – 1). Un second groupe, lié à des différences plus profondes, est constitué de vocables employés dans des commentaires, souvent dépréciatifs, sur les personnages et sur leurs actes : *effronterie* (3 – 2), *gaillard* (3 – 3), *extravagance* (2 – 5), *extravagances* (2 – 1), *sot* (2 – 7), *vice* (2 – 1), *caquet* (1 – 1), *débauche* (1 – 1), *ignorant* (1 – 4), *impertinences* (1 – 2), *incartades* (1 – 1), *lâche* (1 – 3) et *larron* (1 – 1). Cela peut s'expliquer par le ton particulier et par la verve des *Menteurs*. Il s'agit encore, comme dans les premières comédies de Corneille, de chassés-croisés amoureux, mais les propos sont plus libres, et en particulier entre maître et valet, comme c'est généralement le cas dans les comédies à l'espagnole.

5. Mots placé à la rime – spécificités lexicales

Les observations qui suivent sont un échantillon de ce que la statistique lexicale met en évidence par le calcul des spécificités de chaque pièce et de chaque auteur.

[12] Les indications numériques après chaque vocable sont : la fréquence à la rime dans les deux comédies du *Menteur* – la fréquence à la rime dans les 11 pièces de Molière.

Un exemple d'ordre thématique d'abord : les mots-rimes liés au badinage et aux confidences amoureuses sont infiniment plus variés et plus nombreux dans les pièces de Corneille que dans celles de Molière. Les vocables suivants sont des mots-rimes propres à Corneille : *allumer*[13] (9), *dédaigner*[14] (9), *fidélité* (8), *tyrannie*[15] (7), *insensible* (6), *refroidir*[16] (6), *submission/soumission* (6). Les suivants, classés selon leur degré de spécificité décroissante, sont des mots-rimes excédentaires chez Corneille : *maîtresse, affection, amour, flamme, amant, aimer, feu, galanterie* et *volage*. Sur ce point, l'auteur dramatique le plus proche de Corneille pourrait bien être Marivaux.

Pour les registres et les niveaux de langue, on mentionnera, parmi les mots-rimes propres à Molière, des appellatifs et qualificatifs péjoratifs, certains étant parfois employés dans des invectives : *damoiseau*[17] (6 / 9 - 0), *cocu* (3 / 12 - 0), *fripon* (3 / 10 - 0), *cornu* (2 / 2 - 0), *coquin* (2 / 15 - 2), *faquin* (2 / 6 - 0), *jocrisse* (2 / 2 - 0), *pécore* (2 / 2 - 0), *pendard* (2 / 8 - 0), *poltron* (2 / 5 - 3), *vaurien* (2 / 2 - 0), *apôtre* (1 / 1 - 0), *carogne* (1 / 2 - 0), *chattemite* (1 / 1 - 0), *cornard* (1 / 1 - 0), *drôle* (1 / 5 - 0), *étourneau* (1 / 1 - 0), *gueux* (1 / 5 - 0), *hère* (1 / 1 - 0), *péronnelle* (1 / 1 - 0) et *scélérat* (1 / 9 - 1). Ces vocables, qui pourraient être cantonnés dans des farces ou dans des « pièces farcesques » (Rohou, 1996 : 126), sont assez largement attestés dans toutes les pièces comiques, en particulier dans *l'École des femmes*, *Tartuffe* et *les Femmes Savantes*. Plusieurs d'entre eux, tels que *fripon* et *apôtre*, sont attestés à la rime dans *les Plaideurs*.

Jeux de scène et bastonnade. Le mot *bâton*, attesté 5 fois à la rime[18] chez Molière, qui a 17 attestations au total dans les comédies en alexandrins et plus de 60 attestations dans l'ensemble de son théâtre, n'apparaît jamais dans l'édition de 1682 de l'œuvre de Corneille. Il a cependant eu une occurrence dans la première édition de la *Galerie du Palais*, avant que Corneille ne le fasse disparaître, dans l'édition de 1660 : « […] cent *coups de bâton* qu'il reçut l'autre jour » (acte I, scène 9) devient « […] quelques *tours de main* qu'il reçut l'autre jour ». Notons que le mot a deux occurrences dans *les Plaideurs*, mais jamais en tant que mot-rime.

Pour la typologie grammaticale des mots-rimes, on constate, parmi les plus excédentaires chez Molière, un nombre important de mots-outils : *voilà, là, ceci, ainsi aussi, même, rien, ici, pas, cela, peu* (selon l'ordre décroissant des écarts). Il convient d'ajouter que *ceci* (12 attestations à la rime pour Molière) n'apparaît qu'une seule fois à la rime chez Corneille ; *cela* (21 attestations chez Molière) n'apparaît jamais à la rime dans les pièces de Corneille.

B-M Kylander a observé, dans sa thèse, que la longueur moyenne du mot est plus petite dans les comédies de Molière que dans celles de Corneille (Kylander, 1995 : 58). Ce caractère, établi en prenant en compte tout le vocabulaire trouve un écho dans les mots placés à la rime. Ainsi, le calcul de l'écart réduit sur un ensemble de mots monosyllabiques[19] montre un déficit

[13] La fréquence à la rime dans Corneille figure entre parenthèses. Comme tous ces mots pourraient avoir de nombreux autres emplois, nous donnons ci-dessous quelques contextes illustrant le sens qui est noté ici. « Quelques feux dans ton cœur que ton amant *allume*, » *La Suivante* v. 1042.

[14] « S'il m'aime, il se punit en m'osant *dédaigner* » *L'Illusion comique* v. 847.

[15] « L'amour use sur moi de trop de *tyrannie*. » *La Galerie du Palais* v. 281.

[16] « Ne sollicite plus mon âme *refroidie* » *La Place Royale* v. 1334.

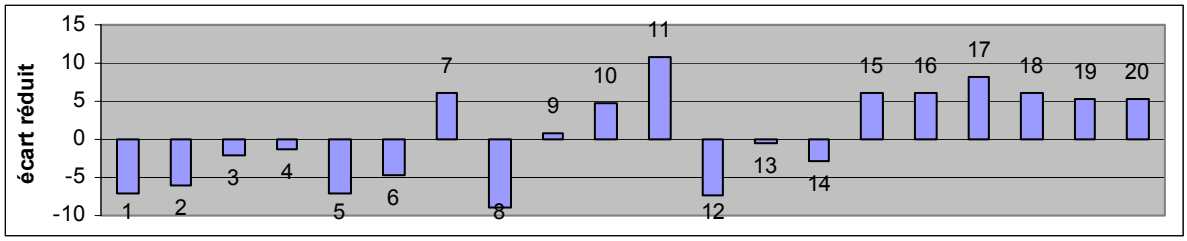
[17] Les indications numériques après chaque vocable sont : la fréquence dans les 11 comédie de Molière à la rime / la fréquence totale dans les mêmes pièces – la fréquence totale dans les 8 comédies de Corneille.

[18] La répartition est la suivante : *l'Étourdi* (2), *le Dépit amoureux* (1), *l'École des femmes* (1) et *Tartuffe* (1).

[19] Cet ensemble comporte les adverbes de négation *pas, point, guère* et *rien* et les pronoms personnels *je, moi, tu, toi, il, elle, lui, soi*, etc.

significatif chez Corneille (écart réduit z = 11,63 correspondant à une probabilité inférieure à 10^{-09}).

L'examen des écarts obtenus pièce par pièce permet de nuancer ce constat : *le Menteur*, qui se distingue aussi sur ce point, est la seule pièce qui soit excédentaire. Parmi les pièces de Molière, seules deux pièces sont manifestement déficitaires : *Dom Garcie* et *les Fâcheux.* Cette dernière est la première comédie-ballet de Molière, commandée par Fouquet pour un spectacle destiné à la Cour.

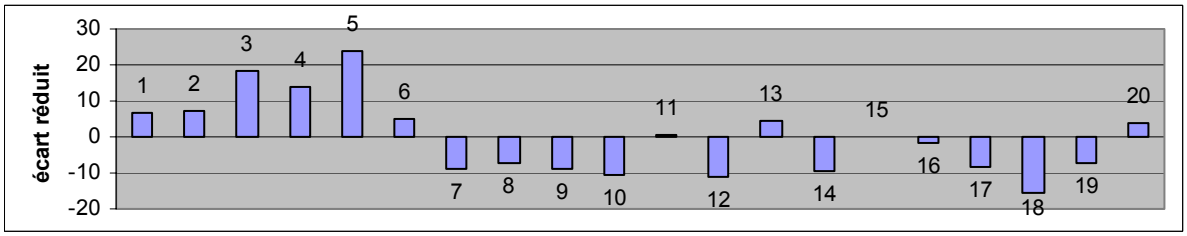


Mots courts[20]

Une dernière remarque notable du point de vue de l'histoire du lexique concerne la répartition du mot *parfois*. Ce mot est 5 fois un mot-rime chez Molière ; il figure 19 fois dans les 11 comédies en vers et 45 fois dans l'ensemble de l'œuvre de Molière. Il a une occurrence dans *les Plaideurs*, mais pas en tant que mot-rime. On ne le trouve jamais dans le théâtre de Corneille. Richelet, dans l'édition augmentée de son dictionnaire de 1693 indique : « Ce mot signifie *quelquefois*, mais il n'est pas si usité que *quelquefois* » ; le *dictionnaire de l'Académie*, en 1694, l'enregistre sans mention particulière mais souligne, au siècle suivant, son caractère familier : « Il n'est guère que du style le plus familier » (éd. de 1798).

6. Quelques rimes caractéristiques

On examinera successivement quelques types de rimes dont les effectifs sont suffisants pour se prêter à un calcul de l'écart réduit.



Rimes en -ion(s)[21]

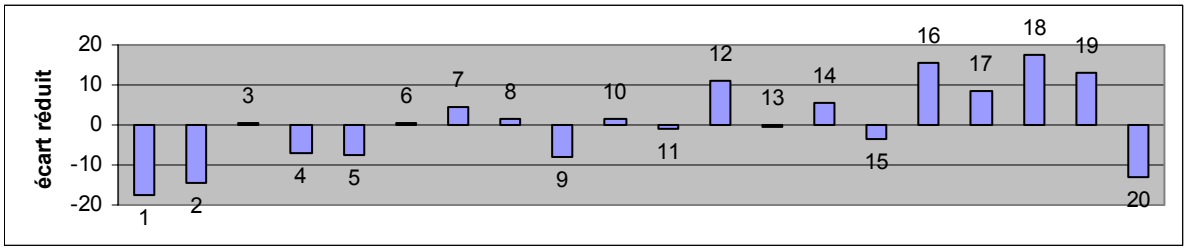
On a observé que les mots en *-ion* « ont tendance à devenir plus rares avec le temps chez Corneille » (Muller, 1967 : 145) et qu'ils sont plus rares dans les tragédies de Racine que dans

[20] Dans ce tableau, ainsi que dans ceux du même type qui suivent, les huit premières pièces sont celles de Corneille, les onze suivantes celles de Molière et la dernière la comédie de Racine. Corneille : *Mélite* (1), *La Veuve* (2), *La Galerie du Palais* (3), *La Suivante* (4), *La Place Royale* (5), *l'Illusion comique* (6), *Le Menteur* (7), *La Suite du Menteur* (8). Molière : *L'Étourdi* (9), *Dépit amoureux* (10), *Sganarelle* (11), *Dom Garcie de Navarre* (12), *l'École des maris* (13), *les Fâcheux* (14), *l'École des femmes* (15), *Le Tartuffe* (16), *Le Misanthrope* (17), *Mélicerte* (18), *les Femmes savantes* (19). Racine : *les Plaideurs* (20). La succession numérique ne s'écarte de la chronologie que sur un point : la comédie des *Plaideurs* est antérieure à celle des *Femmes savantes*.

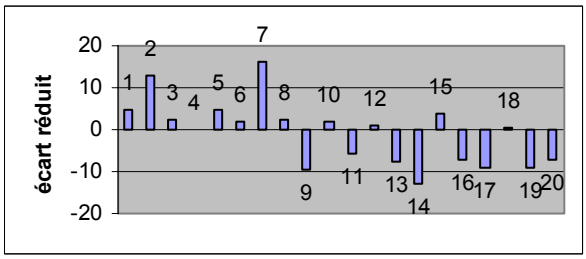
[21] Le test est effectué sur les formes du singulier et du pluriel (*–ion* et *–ions*).

celles de Corneille (Bernet, 1983 : 187-189). L'une des explications possibles tient à la prosodie et à la métrique : dans la versification classique, la séquence *-ion* imposait la diérèse[22], qui ne correspondait probablement plus, au milieu du 17e siècle, à la prononciation commune. Si cette hypothèse correspond à la réalité, Molière, qui était partisan d'une déclamation naturelle[23], aurait dû être tenté d'éviter l'emploi de ces mots. La comparaison de Corneille et Molière sur ce point s'impose. L'excédent de Corneille est hautement significatif (écart réduit z = 9,09 avec une probabilité inférieure à 10^{-9}). L'examen de la suite des pièces montre un très net excédent dans les pièces de Corneille antérieures à 1640, suivi d'une inversion de tendance à partir du *Menteur*. Les pièces de Molière sont très majoritairement déficitaires, les écarts de celles qui ne le sont pas sont peu significatifs. La comédie de Racine donne un écart positif, lui-même peu significatif, qui montre de ce fait une tendance moins affirmée que chez Molière.

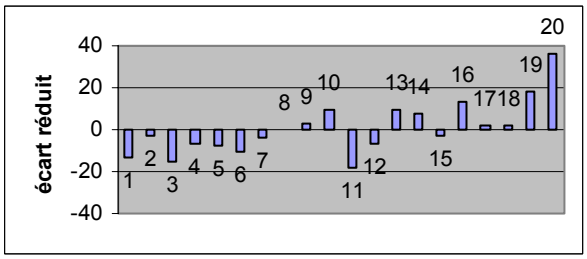
On illustrera quelques-unes des tendances propres à nos auteurs avec la répartition de rimes masculines dont le noyau vocalique est [u] et de rimes féminines dont le noyau vocalique est [ɛ]. Dans le premier cas, le test porte sur des mots à finale graphique *-ous* ou *-oux,* homophones en fin de vers, (ex. : *coups, jaloux* et des mots-outils fréquents comme *nous, vous, tous, dessous*). Dans le second cas, le test porte sur le singulier et le pluriel d'une part de mots en *-ècе(s), -esse(s)* ou *-aisse(s)* (ex. : *laisse, adresse, cesse, expresse, espèce, pièce, Lucrèce, caresses, foiblesses, pièces*) et d'autre part de mots en *-aire(s)* ou *-ère(s)* (ex. : *faire, plaire, altère, entière, père, Valère, affaires, téméraires, prières*).



Rimes -ous/-oux



Rimes -ècе(s)/-esse(s)



Rimes -aire(s)/-ère(s)

Dans tous les graphiques, on constate une tendance dominante pour chaque auteur. Les huit comédies de Corneille manifestent clairement des traits communs pour les rimes à noyau vocalique en [ɛ]. L'unité n'est pas aussi manifeste dans la suite des comédies de Molière.

[22] Sauf dans les terminaisons verbales.

[23] L'« esthétique du naturel » est l'un des caractères affirmés du classicisme (v. Forestier, 1993 : 10-11 ; Rohou, 1996 : 102-103). En s'opposant à la tradition, Molière était, d'une certaine manière, en accord avec l'art de son temps.

On terminera cette brève étude en donnant un exemple, et un seul, d'associations lexicales à la rime. La déclamation en usage au théâtre impliquait la prononciation des consonnes finales[24], y compris pour les infinitifs des verbes du premier groupe[25] ; l'association de *arrêter* avec *Jupiter* (*L'Illusion comique* v. 275-276) ou de *arracher* avec *chair* (*L'Étourdi* v. 1941-1942) n'avait donc rien que de normal. Les rares associations de cette nature figurent, chez Molière, principalement dans ses premières pièces. Lorsque, à la lumière de ces indications, on recense les associations à la rime du mot *air* dans le corpus des pièces comiques, on relève chez Corneille : *air/parler*, *hurler/air* (*Mélite*), *parler/air* (*La Veuv*e), *air/parler* (*La Suivante*), *air/parler* (*L'Illusion comique*), *air/donner, air/parler* (*Le Menteur), air/accorder, air/parler* (*La Suite du Menteur*), et chez Molière : *cher/air* (*L'École des Maris), pair/air* (*les Fâcheux*), *air/chair* (Tartuffe), *air/clair* (*Le Misanthrope*). En d'autres termes, Corneille, en associant toujours le mot *air* avec des infinitifs, montre qu'il adhère pleinement l'usage de cette déclamation, alors que Molière, en ne l'associant qu'avec des mots dont la finale est en toutes circonstances en [ɛR], donne l'impression de s'appuyer plutôt sur la prononciation courante.

*

On trouve, dans un site très sérieux consacré à Molière[26], cette remarque : « les différences existant entre les auteurs [classiques] ne tiennent pas au choix des mots, mais à leur disposition », qui est pour le moins une demi-vérité. Les différences entre Corneille et Molière tiendraient donc simplement à un agencement différent du même stock lexical ; en somme il n'y aurait entre eux que des différences d'ordre syntagmatique ou syntaxique. Nous avons voulu montrer ici que les choix lexicaux pouvaient bel et bien signaler des différences entre les auteurs, même à une époque où les conventions limitaient les manifestations de leur personnalité.

Références

Beaudouin V. (2002). *Mètre et rythmes du vers classique – Corneille et Racine*. Champion.

Beaudouin V. et Yvon Fr. (1996). The Metrometer : a Tool for Analysing French Verse. *Literary and Linguistic Computing*, vol. (11/1), Oxford University Press : 23-32.

Bernet Ch. (1983). *Le Vocabulaire des tragédies de Jean Racine – Étude statistique*. Slatkine-Champion.

Bernet Ch. (1999). Les mots placés à la rime dans le théâtre de Racine. *La Licorne*, vol. (50) : 189-202.

Chaouche S. (2001). *L'Art du comédien – Déclamation et jeu scénique en France à l'âge classique (1629-1680)*. Champion.

Forestier G. (1993). *Introduction à l'analyse des textes classiques*. Nathan.

Gilot M. et Serroy J. (1997). *La Comédie à l'âge classique*. Belin.

Green E. (2001). *La Parole baroque.* Desclée de Brouwer.

Kylander Br.-M. (1995). Le vocabulaire de Molière dans les comédies en alexandrins. *Acta universitatis Gothoburgensis*.

[24] « À la rime, et devant tout autre arrêt de la voix, une consonne finale doit s'articuler. » (Green, 2001 : 284) et « [L'] articulation des consonnes finales annule par la même occasion le mythe [des] rimes normandes » (Chaouche, 2001 : 286).

[25] « Le e qui précède l'r finale, s'ouvrait obligatoirement suivant la loi phonétique qui impose d'ouvrir la voyelle devant une consonne articulée » (Chaouche, 2001 : 286).

[26] http//www.toutmoliere.net ; consulté en novembre 2003.

Labbé D. et Hubert P. (1988). Un modèle de partition du vocabulaire, in Labbé D., Serant D. et Thoiron Ph., *Études sur la richesse et la structure lexicales*. Champion-Slatkine.

Labbé C. et Labbé D. (2001). Inter-Textual Distance and Authorship Attribution Corneille and Molière. *Journal of Quantitative Linguistics*, vol. (8/3) : 213-231.

Lebart S. et Salem A. (1994). *Statistique textuelle.* Dunod.

Muller Ch. (1967). *Étude de statistique lexicale – Le Vocabulaire du théâtre de Pierre Corneille*. Larousse.

Rohou J. (1996). *Le Classicisme*. Hachette.

ANNEXE

			Mots-formes				*Lemmes*					
Pièces – (date de la 1ère repr.)	**N**	**V**	**E(V')**	**lim -**	**lim +**	**E(V'')**	**V**	**E(V')**	**lim -**	**lim +**	**E(V'')**	**Rang**
Scédase (Hardy)	1368	1023	797	768,5	825,5	518						
***L'Étourdi (Molière)* – 1654**	2068	1329	757	723,7	790,3	496	1116	675	645,5	704,5	458	1
Mélite (P. Corneille) – 1629	1822	1190	751	718,9	783,1	492	969	652	624,1	679,9	447	3
Lucrèce (Hardy)	1324	924	743	716,6	769,4	489						
Hector (Montchrestien)	2386	1421	741	708,0	774,0	489						
***L'École des Femmes (Molière)* – 1662**	1779	1145	740	708,6	771,4	489	977	660	631,7	688,3	450	2
Clitandre 1ère éd. (P. Corneille)	1872	1170	736	685,5	744,5	487						
***Fâcheux (Molière)* – 1661**	826	634	---	---	---	486	564	---	---	---	442	5
***Tartuffe (Molière)* – 1664**	1962	1226	733	700,8	765,2	482	1016	643	614,6	671,4	438	6
La Suite du Menteur (P. Corneille) – 1643	1904	1163	725	693,8	756,2	482	960	637	609,6	664,4	439	7
L'Illusion Comique (P. Corneille) – 1636	1688	1074	725	694,6	755,4	480	925	651	623,4	678,6	446	4
***Le Dépit amoureux (Molière)* – 1656**	1796	1117	720	689,1	750,9	477	945	636	608,3	663,7	436	9
Le Menteur (P. Corneille) – 1642	1804	1111	719	688,2	749,8	477	920	631	603,9	658,1	436	10
Athalie (Racine)	1816	1121	717	686,2	747,8	477	927	633	605,9	660,1	438	
Clitandre (P. Corneille)	1624	1019	715	685,5	744,5	477						
Cinna (P. Corneille)	1780	1074	711	680,9	741,1	476						
Polyeucte (P. Corneille)	1814	1092	708	677,7	738,3	473						
Les Plaideurs (Racine) –1668	882	648	---	---	---	472	589	---	---	---	437	8
Phèdre (Racine)	1654	1019	704	674,5	733,5	469	840	615	588,8	641,2	429	
La Veuve (P. Corneille) – 1633	1982	1157	702	671,2	732,8	468	960	619	592,0	646,0	428	12
***Le Misanthrope (Molière)* – 1666**	1808	1076	700	669,9	730,1	467	897	613	586,3	639,7	424	13
Pompée (P. Corneille)	1812	1078	697	666,9	727,1	466						
Médée 1ère éd. (P. Corneille)	1660	1007	696	666,8	725,2	466						
Médée (P. Corneille)	1628	991	695	666,0	724,0	466						
***L'École des Maris (Molière)* – 1661**	1114	754	695	676,8	713,2	466	649	604	586,6	621,4	421	14
***Sganarelle (Molière)* – 1660**	657	501	---	---	---	465	466	---	---	---	435	11
Esther (Racine)	1356	867	692	666,1	717,9	463	732	605	581,0	629,0	423	
La Place Royale (P. Corneille) – 1635	1529	934	690	662,0	718,0	464	779	601	575,8	626,2	422	18
***Les Femmes Savantes (Molière)* – 1672**	1776	1054	688	658,2	717,8	460	893	605	578,1	631,9	417	15
La Galerie du Palais (P. Corneille) – 1634	1794	1037	687	657,6	716,4	463	860	602	576,1	627,9	422	16

La Suivante (P. Corneille) – 1634	1700	1005	685	655,8	714,2	459	837	602	576,1	627,9	420	17
Théodore (P. Corneille)	1882	1050	676	646,7	705,3	457						
Iphigénie (Racine)	1796	1017	673	643,9	702,1	454	827	584	558,9	609,1	413	
La Toison d'Or (P. Corneille)	2237	1179	672	642,2	701,8	454						
Sertorius (P. Corneille)	1920	1059	672	642,7	701,3	454						
Britannicus (Racine)	1768	1001	670	641,1	698,9	453	832	590	564,5	615,5	415	
Bajazet (Racine)	1748	984	664	635,4	692,6	449	791	572	547,3	596,7	406	
***Dom Garcie (Molière)* – 1661**	1867	1003	660	631,6	688,4	450	825	578	553,4	602,6	411	19
Mithridate (Racine)	1698	948	655	626,9	683,1	445	771	567	542,5	591,5	403	
Othon (P. Corneille)	1832	1005	652	623,2	680,8	440						
Attila (P. Corneille)	1788	974	651	622,6	679,4	442						
***Mélicerte (Molière)* – 1666**	600	440	---	---	---	(440)	394	---	---	---	(394)	20
Rodogune (P. Corneille)	1844	1012	651	622,0	680,0	438						
Andromède (P. Corneille)	1772	970	650	621,7	678,3	442						
Nicomède (P. Corneille)	1854	994	650	621,6	678,4	442						
Œdipe (P. Corneille)	2010	1051	648	619,2	676,8	440						
Don Sanche d'Aragon (P. Corneille)	1830	985	646	617,6	674,4	439						
Suréna (P. Corneille)	1738	930	637	609,2	664,8	434						
Pertharite (P. Corneille)	1854	961	634	606,2	661,8	433						
Horace (P. Corneille)	1782	927	628	600,6	655,4	432						
La Mort de Mithridate (La Calprenède)	1758	915	627	599,7	654,3	431						
Sophonisbe (P. Corneille)	1822	938	627	599,5	654,5	430						
Héraclius (P. Corneille)	1916	975	625	597,1	652,9	427						
Le Cid 1ère éd. (P. Corneille)	1866	949	624	596,4	651,6	427						
Agésilas (P. Corneille)	2122	1027	623	595,3	650,7	429						
Le Cid (P. Corneille)	1840	940	619	591,4	646,6	423						
Alcionée (Du Ryer)	1686	856	609	582,7	635,3	424						
Tite et Bérénice (P. Corneille)	1774	902	608	580,8	635,2	416	759	534	509,8	558,2	377	
Tiridate (Campistron)	1308	728	606	582,8	629,2	418						
Pulchérie (P. Corneille)	1758	894	606	579,0	633,0	415						
Andromaque (Racine)	1648	849	604	577,7	630,3	416	691	522	499,0	545,0	377	
Le Comte d'Essex (La Calprenède)	1720	847	598	571,9	624,1	415						
Bérénice (Racine)	1506	776	594	568,9	619,1	414	659	522	499,2	544,8	376	
Ariane (Th. Corneille)	1754	839	580	554,0	606,0	400						
La Thébaïde (Racine)	1516	759	579	554,3	603,7	405	618	493	471,2	514,8	360	
Alexandre (Racine)	1548	767	575	550,1	599,9	400	634	497	474,9	519,1	361	
Astrate (Quinault)	1662	805	574	548,5	599,5	399						
Phèdre et Hippolyte (Pradon)	1738	746	533	508,9	557,1	377	644	478	456,3	499,7	348	
Armide (Quinault)	830	444	---	---	---	362						

Tableau 1. Mots-formes et lemmes à la rime – Ordre décroissant

	La Galerie du Palais		**Le Menteur**		**La Suite du Menteur**		***L'École des Maris***		***L'École des Femmes***		***Les Femmes Savantes***	
	formes	lemmes	formes	lemmes	formes	lemmes	formes	lemmes	formes	lemmes	formes	lemmes
Mélite	0,635	0,547	**0,729**	**0,635**	**0,709**	**0,626**	0,734	0,655	0,754	0,688	0,768	0,696
La Veuve	0,625	0,542	**0,702**	**0,627**	**0,701**	**0,621**	0,711	0,637	0,746	0,677	0,729	0,676
La Galerie du Palais	0	0	**0,703**	**0,625**	**0,700**	**0,612**	0,713	0,644	0,741	0,675	0,738	0,667
La Suivante	0,606	0,513	**0,720**	**0,637**	**0,731**	**0,634**	0,737	0,664	0,761	0,691	0,742	0,681
La Place Royale	0,601	0,514	**0,723**	**0,642**	**0,708**	**0,629**	0,743	0,672	0,757	0,690	0,758	0,695
L'Illusion Comique	0,663	0,577	0,695	0,629	0,686	0,600	0,733	0,668	0,735	0,666	0,733	0,684
Le Menteur	0,703	0,625	0	0	0,670	0,569	0,727	0,635	0,728	0,646	0,733	0,651
La Suite du Menteur	0,700	0,612	0,670	0,569	0	0	0,700	0,619	0,714	0,634	0,730	0,652
L'Étourdi	0,704	0,610	0,699	0,615	0,702	0,617	0,687	0,602	0,677	0,595	0,716	0,636
Le Dépit amoureux	0,678	0,605	0,704	0,630	0,697	0,609	0,661	0,586	0,652	0,577	0,696	0,608
Dom Garcie	0,674	0,598	0,734	0,659	0,710	0,633	0,682	0,603	**0,737**	**0,658**	**0,716**	**0,646**
L'École des Maris	0,713	0,644	0,727	0,635	0,700	0,619	0	0	0,690	0,619	0,688	0,588
L'École des Femmes	0,742	0,675	0,728	0,646	0,714	0,634	0,690	0,619	0	0	0,698	0,625
Tartuffe	0,730	0,654	0,720	0,644	0,715	0,629	0,646	0,560	0,657	0,564	0,652	0,568
Le Misanthrope	0,684	0,617	0,721	0,646	0,681	0,601	0,664	0,594	0,667	0,598	0,625	0,553
Les Femmes Savantes	0,738	0,667	0,733	0,652	0,730	0,653	0,688	0,588	0,698	0,625	0	0

Tableau 2. Distance intertextuelle (Calcul à partir des effectifs des formes / des lemmes – Extrait)

Classification et désarticulation de graphes de termes

Anne Berry[1], Bangaly Kaba[1], Mohamed Nadif[2], Eric SanJuan[2],
Alain Sigayret[1]

[1] LIMOS – Université Blaise Pascal – France
[2] IUT de Metz, LITA – Université de Metz – France
{berry, kaba, sigayret}@isima.fr, {nadif, eric.sanjuan}@iut.univ-metz.fr

Abstract

Algorithm CPLCL (Classification by Preferential Clustered Link) for the classification graphs of syntactic variations, introduced by Ibekwe-SanJuan (1997), has the mathematical properties of single link clustering, since it is partly founded on the concept of component related to a subgraph, while avoiding the chain effect as much as possible. In spite of these properties, this algorithm of hierarchical classification does not make it possible to avoid the formation of some classes of too great dimension. However, the visualization of subgraphs corresponding to these classes with the AiSee interface highlights obvious structural graph properties, which enable their decomposition with recent formal tools. This motivates us in developing underlying graph-theoretical aspects with algorithm CPCL presented in TermWatch and in defining a decomposition algorithm for these classes.

Résumé

Ibekwe-SanJuan (1997) introduit un algorithme de classification de graphes de relations syntaxiques, qui possède les propriétés mathématiques de la classification par lien simple (CLS), tout en limitant les effets de chaînes. Malgré cela, cet algorithme de classification hiérarchique, implanté dans le système TermWatch, ne permet pas d'éviter la formation de quelques classes de trop grande dimension. Cependant, la visualisation des sous-graphes correspondant à ces classes avec l'interface AiSee met en relief d'évidentes propriétés structurelles de graphes qui permettent leur décomposition avec des outils formels récents. Le développement des aspects de la théorie des graphes sous-jacents à l'algorithme CPCL nous a alors permis d'aboutir à un algorithme de décomposition que nous introduisons ici.

Keywords: text mining, clustering, ultrametrics, minimal separators, terminology, linguistic relations, graph decomposition.

1. Introduction

La principale motivation de cette communication est d'illustrer comment la théorie des graphes peut aider à concevoir des méthodes de classification de données textuelles non fondées sur le seul paradigme de la co-occurrence. En effet, le formalisme de cette théorie permet une représentation efficace d'indices de similarité, avec d'importantes applications en sciences sociales avec la théorie des réseaux sociaux (Social Networks). Les récents outils de visualisation de graphes tels que AiSee[1], offrent de plus des interfaces conviviales et interactives pour explorer les données ainsi représentées. Enfin, et c'est en cela que consiste la principale contribution de cette communication, de récents résultats de cette théorie sur la notion de séparateur minimal permettent de désarticuler des graphes complexes sur leur seules propriétés structurelles.

[1] Cette interface implémente les algorithmes introduits par Sander (1996) et est disponible sur http://www.aisee.com.

On applique ici ces notions à une méthode présentée par Ibekwe-SanJuan et SanJuan (2002a) et qui a été aujourd'hui totalement implantée dans le système TermWatch (Ibekwe-SanJuan et SanJuan, 2002b). Il s'agit d'une méthode de classification non supervisée de termes extraits avec INTEX (Siberztein, 1993), dont le critère d'association est exclusivement fondé sur les relations de variations linguistiques.

Les résultats présentés par Ibekwe-SanJuan et SanJuan (2002a) et validés par un spécialiste en VST (Ibekwe-SanJuan et Dubois, 2002), portaient sur un corpus de textes composé de publications scientifiques en anglais sur les procédés de panification[2], collecté pour répondre à un besoin de veille scientifique et technique (VST).

Le système utilise un algorithme de classification de graphes de relations syntaxiques initialement introduit par Ibekwe-SanJuan (1997) sous le nom de CPCL (Classification by Preferential Clustered Link). Cet algorithme possède les propriétés de la classification par lien simple (CLS), tout en limitant les effets de chaînes sur ce type de données textuelles. Malgré cela, il ne permet pas d'éviter la formation de quelques classes de trop grande dimension, or la visualisation des sous-graphes correspondant à ces classes avec l'interface AiSee met en relief des propriétés structurelles de graphes qui permettent leur désarticulation. Cette observation nous amène à introduire ici un algorithme de décomposition de ces classes dont nous analysons la complexité.

Le reste de l'article est structuré de la façon suivante :

Dans la section 2, nous revenons sur l'extraction par le logiciel INTEX des unités textuelles dont il est question ici et sur leur mise en relation par TermWatch. Nous le faisons en nous basant sur la représentation informatique de ces termes qu'utilise ce système, ce qui nous permet de donner une définition simple du graphe étudié.
La section 3 revient sur l'algorithme CPCL qui est cette fois décrit en termes de réduction de graphes. Ce point de vue nous permet de situer exactement l'algorithme CPCL vis-à-vis de la CLS classique.
La section 4 représente la principale contribution de cet article. Après un exposé des méthodes récentes de désarticulation d'un graphe, nous appliquons cette approche à la décomposition automatique de la plus grosse des classes formées par l'algorithme CPCL sur le corpus de panification. Les propriétés structurelles que cette classe partage avec les autres classes de dimension similaire nous permettent de proposer un algorithme de décomposition de ces classes.

2. Définition d'un graphe de termes

Le but de cette section est de préciser exactement le type de graphe de données textuelles que nous étudions ici. Comme il s'agit d'un graphe automatiquement extrait et réduit par le système TermWatch, il nous est nécessaire de rappeler très sommairement l'architecture de ce système.

TermWatch (Ibekwe-SanJuan et SanJuan, 2002b) est donc un système de classification non supervisée de termes extraits de textes, destiné à la Veille Scientifique et Technique (VST). Appliqué à des ensembles de textes scientifiques et techniques ce système produit une carte de thématiques présentée sous forme d'un graphe. La figure 1 donne l'architecture logicielle de ce système, qui comprend aujourd'hui quatre modules.

2.1. *Extraction des sommets du graphe : les termes*

La technique d'extraction des unités textuelles repose sur des avancées récentes de la terminologie computationnelle (Jacquemin, 2001). Dans le cas du système TermWatch, il s'agit de syn-

[2] Ce corpus a été gracieusement fourni par l'Unité de Recherche et Innovation de l'INIST.

1.	**Extraction de termes d'un corpus de textes en anglais** : appel au logiciel INTEX en mode ligne de commande en utilisant les automates introduits par Ibekwe-SanJuan et SanJuan (2002a).
2.	**Construction du graphe des termes** : Mise en relation des termes par la recherche de variations syntaxiques en utilisant les ressources linguistiques d'INTEX (Dictionnaire DELAF).
3.	**Classification ascendante hiérarchique CPCL** en quatre phases : (a) Partition de l'ensemble des variations en deux classes COMP et CLAS. (b) Extraction des composantes connexes du sous-graphe des variations dans COMP. (c) Réduction du graphe en un graphe valué de l'activité des variations dans CLAS. (d) Agglomération des composantes connexes en classes.
4.	**Génération d'interfaces AiSee et HTML** : - Edition des graphes au format GDL (Graph Description Language) pour AiSee. - Génération de liens hypertextes pour la navigation dans le graphe de termes.

Figure 1. Architecture de TermWatch

tagmes nominaux de plusieurs mots susceptibles de désigner, par leurs propriétés syntaxiques et grammaticales, un objet ou une notion du domaine (donc un terme). Ils sont spécifiques au domaine considéré et ils ont fréquemment une occurrence unique dans tout le corpus de textes.

Comme indiqué précédemment, nous reprenons ici le corpus sur les procédés de panification utilisé dans Ibekwe-SanJuan et SanJuan (2002a). De ces textes ont été extraits 3.651 termes ayant au moins un modifieur, après élimination des termes candidats les plus invraisemblables par un indexeur humain de l'INIST.

Dans le système TermWatch, ces termes candidats sont actuellement codés sous forme de couples modifieurs-centre (M, c) où c représente le *centre* d'un terme et M est une suite de *modifieurs* (essentiellement des noms et des adjectifs) associée à ce centre.

La table 1 montre trois exemples de termes extraits avec leur codage dans le système.

candidat terme	M	c
wheat dough surface stickiness	wheat dough surface	stickiness
baking property of frozen wheat dough flour	frozen wheat dough flour baking	property
greater intensity of aroma	greater aroma	intensity

Table 1. Exemples de termes extraits avec leur représentation dans TermWatch

2.2. *Extraction des arêtes du graphe : les variations syntaxiques*

Dans Ibekwe-SanJuan (1997), il est montré que la notion d'association entre termes, dans un but de classification non supervisée, pouvait être entièrement fondée sur les relations de variations linguistiques que peuvent partager ces termes. Cela représente une intéressante alternative à l'utilisation de la co-occurrence comme critère d'association, tout particulièrement lorsqu'il s'agit d'extraire une information rare.

Pour cette approche, des relations de variation ont été expérimentées ; elles sont décrites en détail dans Ibekwe-SanJuan et SanJuan (2002b)[3].

[3] Cet ensemble de variations, défini sur des opérations de surface d'insertion et de substitution d'éléments, est extrêmement réduit vis-à-vis de celles plus structurelles que peut extraire un système tel que FASTR (Jacquemin, 2001). De plus, le système TermWatch se limite actuellement à la recherche des seules relations syntaxiques, il n'intègre pas la recherche de variations sémantiques telles que la synonymie. L'objectif du système n'étant cepen-

La représentation des termes, précisée dans la sous-section précédente, nous amène à présenter une définition des relations sur lesquelles nous nous baserons pour décrire complètement le graphe qui résulte de la recherche de variations entre termes, ainsi que pour en dégager les premières propriétés structurelles. On considère classiquement quatre types de relations syntaxiques entre termes. Ces relations, utilisées depuis Ibekwe-SanJuan (1997) dans un cadre de classification non supervisée, peuvent être définies comme suit : deux termes $t_1 = (M_1, c_1)$ et $t_2 = (M_2, c_2)$ peuvent être en relation par :

- **COMP** s'ils ont même centre et que M_2 peut être obtenue, à partir de M_1, par substitution d'un unique élément ou par insertion, à une position donnée de M_1, d'une suite de modifieurs.
- **SubCen2** si leurs centres sont différents mais que M_1 et M_2 sont formées d'un même unique élément.
- **SubCen3** si leurs centres sont différents mais que M_1 et M_2 sont composées d'une même suite d'au moins deux mots.
- **Exp** si leurs centres sont différents et que la concaténation de M_1 avec c_1 est une sous-chaîne de M_2 (M_2 est de la forme AM_1c_1B).

Avant de définir le graphe de termes sous-jacent à ces relations de variations, nous allons rappeler quelques notations et définitions fondamentales.

Un **graphe non-orienté** G est un couple (V, E) où V est un ensemble fini quelconque dont les éléments sont appelés **sommets** et E est un ensemble de paires de V (parties de V ayant exactement deux éléments) dont les éléments sont appelés **arêtes**. Une arête e est dite **incidente** à un sommet v si $v \in e$.

A toute relation binaire R sur un ensemble X, on peut associer le graphe non orienté $G_R = (X, E_R)$ où E_R dénote l'ensemble des paires d'éléments de X en relation par R (*i.e.* $E_R = \{\{u, v\} : (u, v) \in R\}$). Réciproquement, à tout graphe $G = (V, E)$, on peut associer la relation binaire symétrique R_G définie sur V par $R_G = \{(v, v) : v \in V\} \cup \{(u, v), (v, u) : \{u, v\} \in E\}$.

Un graphe est une **clique** quand tous les sommets sont deux à deux reliés par des arêtes (*i.e.* $\forall u, v \in V$, $\{u, v\} \in E$). $G' = (V', E')$ est un **sous-graphe** de $G = (V, E)$ si $V' \subseteq V$ et $E' \subseteq E$. Un graphe est dit **connexe** s'il est possible de passer d'un sommet quelconque à un autre par une suite d'arêtes. Un **cycle** est un 'circuit' qui permet de partir d'un point et d'y revenir, sans rencontrer d'arête qui puisse servir de raccourci (**corde**). Un **arbre** est un graphe connexe et sans cycle ; une **forêt** est formée de plusieurs composantes connexes qui sont des arbres.

Il existe des graphes, très proches des arbres, qui forment la classe des *graphes triangulés* ; ils sont définis par l'absence de *cycle* de plus de trois sommets : cela revient à dire que les seuls cycles sont formés par des 'triangles'. La figure 2 donne un graphe triangulé.

Nous pouvons maintenant revenir à notre étude des quatre relations de variations définies ci-dessus. Celles-ci induisent dès lors, sur un ensemble T de termes, un graphe non-orienté $G_0 = G_{COMP \cup CLAS}$ avec CLAS=SubCen2 $\cup$ SubCen3 $\cup$ Exp. Ce graphe est appelé **graphe de termes** de T.

On remarque notamment que :

dant pas l'extraction de ressources terminologiques mais la classification non supervisée de données textuelles, cet ensemble de variations est considéré comme suffisant pour intéresser un veilleur qui cherche à disposer de cartes thématiques (Ibekwe-SanJuan et Dubois, 2002) générées automatiquement. L'adjonction de nouvelles relations pour affiner le processus de classification est à l'étude.

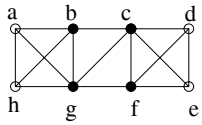


Figure 2. Un graphe triangulé : les seuls cycles sont des 'triangles'

1. Si δ est la longueur maximale d'un terme dans l'ensemble T, $|COMP \cup CLAS| < \delta.|T| << |T|^2$. La matrice d'adjacence du graphe associé à T est très creuse.

2. G_{Exp} est une forêt et $G_{Exp \cup SubCen2}$ est un graphe triangulé.

La remarque 1 justifie l'introduction en Ibekwe-SanJuan (1997) de l'algorithme CPCL spécifique à ce type de données. Nous ré-écrirons cet algorithme dans la section suivante en adoptant le point de vue de la théorie des graphes, ce qui mettra en valeur la compatibilité de cet algorithme avec la conservation de sous-graphes triangulés induits par la remarque 2 précédente. C'est en effet la persistances de ces structures triangulées dans les graphes obtenus par réductions successives, qui est la clef de l'algorithme de désarticulation des graphes volumineux qui sera présenté dans la section 4.

3. Réduction du graphe de termes

Le graphe auquel nous allons appliquer des algorithmes de désarticulation n'est pas le graphes des termes, mais un graphe réduit. Pour parfaitement le décrire il nous faut encore préciser exactement en termes de graphes l'algorithme de classification CPCL mis en œuvre par Term-Watch. On obtient ainsi un exposé simplifié de cet algorithme qui met en relief ses différences et ses points communs avec la CLS classique, ce qui nous permet d'évaluer avec exactitude sa complexité dans le pire des cas.

L'algorithme CPCL introduit en Ibekwe-SanJuan (1997) utilise, d'une part, la relation COMP pour réaliser une première réduction du graphe G_0 et d'autre part, l'union CLAS des autres relations pour dégager des classes susceptibles de représenter des thématiques du domaine au moyen d'une classification ascendante hiérarchique (CAH).

3.1. *Calcul des composantes connexes*

Nous commençons par introduire les notations nécessaires relatives aux graphes réduits.

Si θ est une relation d'équivalence sur V, on note $\theta(V)$ l'ensemble des classes d'équivalence de θ, et pour chaque v dans V, $\theta(v)$ l'élément de $\theta(V)$ contenant v. Si R est une relation binaire quelconque, on note R^* la plus petite relation d'équivalence contenant R. Un graphe $G = (V, E)$ a pour composantes connexes l'ensemble $R^*_G(V)$. Enfin, un sous-graphe $G' = (V', E')$ de G est connexe si $R^*_{G'} = V' \times V'$.

Soit $G = (V, E)$ un graphe et θ une relation d'équivalence sur G. On note $G/_\theta$ le graphe $(\theta(V), E/_\theta)$ tel que $E/_\theta = \{\{\alpha, \beta\} : (\exists u \in \alpha)(\exists v \in \beta)\ \{u, v\} \in E\}$.

Le graphe réduit a comme sommets l'ensemble des classes d'équivalence de θ et l'on trace une arête entre deux classes si et seulement si il existe au moins une arête du graphe initial qui intersecte ces deux classes.

Pour un ensemble de termes, la première étape de la réduction consiste alors à calculer le graphe : $G_0 = G_{COMP \cup CLAS}/_{COMP}$.

3.2. *Agglomération CPCL des composantes connexes*

Pour former des classes susceptibles de représenter des thématiques, on procède alors par classification hiérarchique en choisissant comme critère la somme des proportions des liens de variation de même type entre deux composantes connexes.

Plus précisément, on munit le graphe G_0 d'une valuation d_0 de ses arêtes définie pour tout $\{\alpha, \beta\}$ dans $E_{COMP \cup CLAS}/_{COMP}$ par :

$$d_0(\alpha, \beta) = \sum \left\{ \frac{|(\alpha \times \beta \cup \beta \times \alpha) \cap R|}{|R|} : R \in \{SubCen2, SubCen3, Exp\} \right\}$$

On obtient ainsi un critère de similarité qui tient compte de la nature des liens entre deux agrégats, mais la matrice de similarité induite par ce critère reste une matrice très creuse. Pour réduire G_0 en se basant sur ce critère, il est alors naturel de considérer la classification par lien simple (CLS). En effet, la CLS présente l'avantage d'induire une unique ultramétrique et s'exprime en termes de graphes réduits puisqu'elle consiste à calculer pour les différentes valeurs σ prises par d_0, les graphes G_0/R^*_σ tels que $R_\sigma = \{(\alpha, \beta) : d_0(\alpha, \beta) > \sigma\}$. Cependant, les résultats ainsi obtenus restent insatisfaisants, comme le montre le tableau 2.

La solution présentée par Ibekwe-SanJuan (1997) consiste à confondre les sommets du graphe liés par des arêtes dont la valuation est supérieure à celles des autres arêtes adjacentes. Il s'agit en fait de considérer les maximaux locaux de la fonction d_0. On calcule ainsi une suite de graphes G_i et de valuations d_i telles que pour $i \geq 0$: $G_{i+1} = G_i/_{\theta^*_i}$ avec :

$$\theta_i = \{(\alpha, \beta) : (\forall \gamma)\ (d_i(\alpha, \beta) \geq \max\{d_i(\alpha, \gamma), d_i(\gamma, \beta)\}\}$$

la valuation d_{i+1} étant définie comme dans la CLS par :

$$d_{i+1}(a, b) = \max\{d_i(\alpha, \beta) : \alpha \in a, \beta \in b\}$$

Il découle de cette caractérisation que l'ultramétrique associée à cette classification hiérarchique est unique et inférieure à l'ultramétrique de la CLS. De plus, elle peut être calculée en temps $O(|V|.\delta.|\Im(d)|)$ où $|V|$ est l'ensemble des sommets de G_0, $\delta < |V|$ est le degré maximal d'un sommet $x \in V$ et $|\Im(d)|$ est le nombre de valeurs prises par d.

3.3. *Comparaison des classifications CPCL et CLS des composantes connexes*

L'algorithme présenté dans Ibekwe-SanJuan (1997) semble éviter l'effet de chaîne dans le cas de données creuses discrètes présentant une grande amplitude de valeurs. Sur les données du corpus de panification, on obtient les résultats présentés dans la table 2. Dans cette table, i désigne l'itération, NB_CLS (resp. NB_CPCL) le nombre de classes non triviales (contenant au moins deux éléments) pour la CLS (resp. CPCL), COUV_CLS (resp. COUV_CPCL) le nombre de sommets classés et Max_CLS (resp. Max_CPCL) la taille maximale d'une classe pour la CLS (resp. CPCL).

On constate que le nombre de classes décroît avec i pour l'algorithme CPCL tandis qu'il atteint un maximum à $i = 19$ pour la CLS. C'est la classification associée à $i = 3$ qui représente le meilleur compromis entre le nombre de classes non triviales, leur taille maximale et leur couverture pour l'algorithme CPCL. C'est cette classification qui a été validée dans Ibekwe-SanJuan et Dubois (2002). Il est intéressant de noter qu'aucun niveau de la CLS n'optimise de cette manière ces trois critères.

i	NB_CLS	NB_CPCL	COUV_CLS	COUV_CPCL	Max_CLS	Max_CPCL
1	1	54	3	173	3	9
2	2	44	6	192	3	12
3	3	33	9	195	3	29
7	13	15	48	219	7	176
14	19	13	87	223	29	187
25	13	13	190	223	125	187

Table 2. Nombre, couverture et taille maximale des classes obtenues par les algorithmes CLS et CPCL lors de différentes itérations

4. Une décomposition structurelle des clusters

Quoique la méthode de classification précédente limite les effets de chaîne, certaines classes restent trop grosses pour être interprétables par l'utilisateur. Ainsi, dans le cas du corpus sur la panification, la plus grosse classe obtenue à la troisième itération avec l'algorithme CPCL contient 219 termes répartis en 29 composantes. La structure de cette classe automatiquement libellée *dough behaviour* est développée dans la figure 3 telle que AiSee permet de la visualiser avec ses liens vers les autres classes. Les éléments de cette classe *dough behaviour* sont représentés par des rectangles et sont disposés vis-à-vis des autres classes représentées par des cercles. Comme cela a été signalé par Ibekwe-SanJuan et SanJuan (2002b), on trouve dans cette classe deux grosses composantes (n° 1, 2) très proches, formées autour des termes *wheat protein* et *wheat flour dough* qui totalisent l'essentiel des liens vers l'extérieur (vers les autres classes et composantes). On remarque aussi que ce sont deux très petites composantes (n° 3, 4) formées autour des noms-centres *fibre* (*cellulose fibre, dietary fibre,...*) et *roll* (*crisp roll, hard roll, bread roll*) qui jouent un rôle de "passerelle" ou de "connecteur" dans la structure de cette classe.

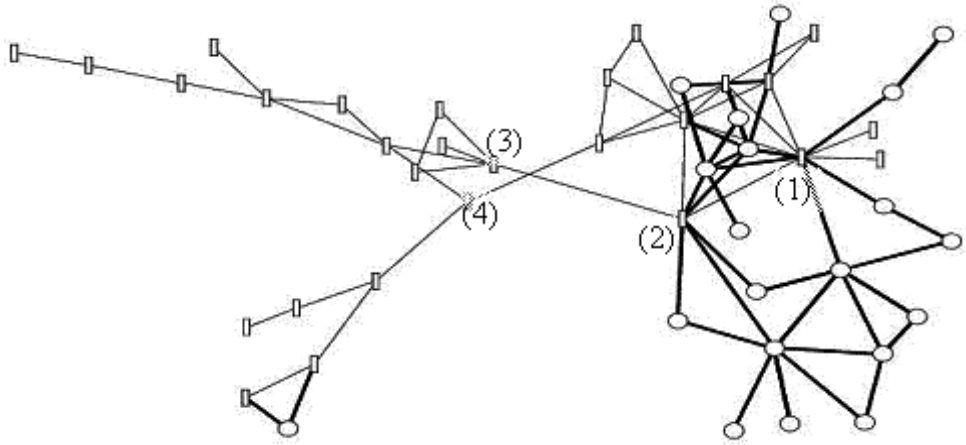


Figure 3. Structure de la classe dough behaviour

Nous allons dans cette section nous attacher à décomposer ce type de classe en nous efforçant de respecter leur cohérence structurelle.

4.1. *Les séparateurs minimaux complets*

Des travaux récents ont montré que les graphes pouvaient se décomposer de façon très cohérente en utilisant leurs articulations naturelles, que nous allons décrire, et qui sont basées exclusive-

ment sur des propriétés structurelles des graphes. Des premiers résultats très intéressants ont été obtenus avec l'examen des concepts engendrés par une base de données binaire, dans la recherche systématique de motifs pour la fouille de données (data mining) (Berry et Sigayret, 2002).

Dans les graphes simples que sont les arbres, les articulations sont les noeuds qui ne sont pas des feuilles, les feuilles étant les sommets de l'arborescence qui ont au plus une arête incidente. Le propre d'une articulation est que son retrait définit plusieurs sous-graphes connexes. De bonnes décompositions sont obtenues en recopiant une articulation dans les différents sous-graphes ainsi définis.

Lorsque le graphe considéré n'est pas un arbre, comme c'est le cas du graphe représentatif de notre corpus, on se ramène utilement à cette décomposition en cherchant des articulations qui ne sont pas formées d'un seul sommet, mais d'un groupe de sommets qui sont tous reliés entre eux. On exige toujours bien sûr, que le retrait de ce groupe de sommets définisse plusieurs sous-graphes connexes.

Pour préciser les outils que nous mettons en œuvre, sur un graphe qui n'est pas un arbre, les articulations que nous utilisons portent l'appellation de *séparateurs minimaux*. Nous pouvons en donner la définition formelle suivante : un sous-ensemble S de sommets est un *séparateur minimal* si en retirant les sommets de S ainsi que toutes les arêtes issues des sommets de S, on obtient un graphe qui n'est pas connexe (c'est-à-dire que l'on ne peut pas forcément aller d'un point à un autre en suivant un chemin dans le graphe); de plus, on demande qu'il y ait deux sommets a et b qui sont séparés par le retrait de S (a et b appartiennent à deux parties connexes ainsi définies), et tels qu'aucun sous-ensemble propre de S ne parvienne à séparer a de b par son retrait.

Un graphe, en général, peut posséder un nombre exponentiel de séparateurs minimaux, ce qui rend cet outil inimplémentable ; par contre, il est prouvé que le nombre de séparateurs minimaux complets (*i.e.* formant une clique) est faible (il y en a moins que de sommets). De plus, la décomposition décrite ci-dessus pour les arbres, et qui s'appelle en général décomposition par *séparateurs minimaux complets*, préserve les cycles (Trajan, 1985), et la décomposition du graphe ainsi que l'énumération des tous les séparateurs minimaux complets peut se faire rapidement (en temps $O(|V||E|)$, voir Berry et Bordat, 1997), à l'aide d'algorithmes tels que LEX M (Rose *et al.*, 1976) ou MCS-M (Berry *et al.*, 2002).

Un des critères qui permet de dire que les graphes triangulés sont très proches des arbres est que tous leurs séparateurs minimaux sont complets, et qu'il y en a un très petit nombre, ce qui les fait ressembler aux points d'articulation d'un arbre.

Une décomposition par séparateurs minimaux complets peut se faire soit entièrement, en recommençant sur les sous-graphes obtenus jusqu'à ce qu'il ne reste plus de séparateur complet dans aucun d'entre eux,

4.2. *Application à la classe* Dough behaviour

Examinons maintenant la classe *Dough behaviour* de notre corpus, visualisé sous forme d'un graphe. A quelques anomalies près, la structure présentée est très proche de celle d'un arbre. Il apparaît seulement quelques anomalies : d'abord quelques cycles de longueur trois (des 'triangles'), qui sont des anomalies qui ne la font pas s'éloigner de façon significative d'un arbre, comme nous l'avons expliqué ci-dessus; ensuite, il apparaît un unique cycle de longueur six, représenté en gras sur la figure 4.

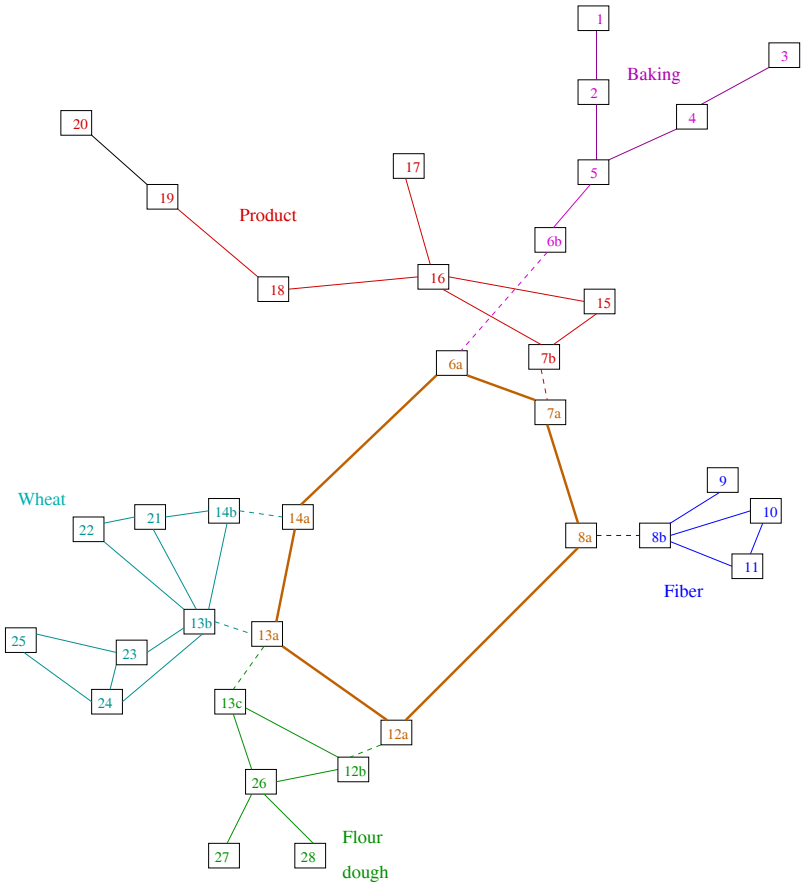


Figure 4. Une décomposition de la classe Dough behaviour *suivant des articulations équilibrantes ; les pointillés représentent les endroits où l'on a recopié des sommets d'une articulation pour décomposer*

Notre première impression est que, étant donnée la structure presque arborescente du graphe, le cycle représente réellement une anomalie; nous avons donc tenté de l'utiliser comme base pour notre décomposition. Nous avons en conséquence tout d'abord choisi des articulations qui bordaient cette anomalie, ce qui permet de l'isoler. Nous obtenons, en plus de ce cycle, deux morceaux qui sont des arbres, et trois morceaux avec quelques triangles, mais qui restent très proches d'un arbre puisqu'ils sont des graphes triangulés. Le résultat est relativement bien équilibré et les 6 morceaux obtenus pas trop gros. La figure 4 illustre ce résultat. Nous y avons renommé les sommets, pour plus de simplicité. Un premier intérêt de la notion de séparateur minimal est donc de pouvoir visualiser un sous-graphe aussi imbriqué que celui des sommets rectangulaires de la figure 3 par un ensemble d'articulations représentées en figure 4.

Cependant, pour vérifier de manière empirique si la décomposition obtenue respectait la nature syntaxique des liens représentés par ce graphe, nous avons recherché des éléments lexicaux récurrents dans les libellés des composantes appartenant à une même articulation. Ces libellés sont les termes de la composante ayant un nombre maximal de variantes (Termes avec la plus forte activité de variation). Par définition, il se peut que cet élément ne soit pas unique. Nous avons alors trouvé un élément léxical qui caractérise chacun des cinq morceaux différents du cycle. Ces éléments sont indiqués sur la figure 4. Cela signifie qu'une composante appartient à l'une de ces cinq articulations si et seulement si au moins un de ses libellés contient l'élément lexical associé à l'articulation. Il est possible de trouver d'autres décompositions de cette classe

avec cette même propriété, l'intérêt de la décomposition proposée ici est d'avoir pu être déduite des seules propriétés structurelles du graphe. D'autant plus que l'on obtient cette même propriété des éléments lexicaux pour la décomposition des quatre autres plus grosses classes qui présentent respectivement :

– Une structure d'arbre (sans cycle) ;
– Un cycle de longueur 4 plus un cycle de longueur 3 ;
– Un cycle de longueur 3 ;
– Trois cycles de longueur 3 et un cycle de longueur 4.

Il est cependant aisé de trouver des contre-exemples théoriques pour lesquels la décomposition introduite ici n'aura pas cette propriété des éléments lexicaux récurrents, d'où la nécessité de confronter cette décomposition à un grand nombre de corpus pour la valider de manière empirique. Ce travail est en cours.

4.3. *Algorithme de décomposition*

Nous proposons de façon plus générale et systématique le procédé algorithmique suivant, qui dans un premier temps décompose complètement le graphe, ce qui permet de récupérer d'une part la liste des séparateurs minimaux complets et d'autre part la liste des sous-graphes correspondants (appelés des 'atomes'); dans un deuxième temps, on examine les atomes contenant des cycles anormaux (c'est-a-dire de longueur au moins 4) et on utilise les séparateurs complets qui les bordent pour décomposer le graphe de départ. Après cette phase, si l'un des sous-graphes obtenus reste trop gros, on le redécompose en utilisant un séparateur complet qui est 'central'.

Pour décomposer, on utilise une numérotation des sommets fournie par LEX M (Rose *et al.*, 1976) ou MCS-M (Berry *et al.*, 2002), ce qui coûte $O(|V||E|)$, suivi du procédé proposé par Trajan (1985) et détaillé dans Berry et Bordat (1997), qui ne requiert pas de temps supplémentaire. Déterminer si un atome ne contient pas de cycle de longueur au moins 4 revient à vérifier si cet atome est un graphe triangulé, ce qui coûte $O(|E|)$, en utilisant un autre algorithme (Rose *et al.*, 1976). Enfin, trouver un séparateur minimal complet qui soit 'central' peut se faire en utilisant la numérotation calculée par LEX M ou MCS-M puis en prenant un numéro 'moyen'; cette approximation marche bien dans la pratique. On obtient l'algorithme suivant :

Algorithme de décomposition de graphes de termes

Donnée : Un graphe $G = (V, E)$.

Résultat : Une décomposition de G à l'aide de séparateurs minimaux complets.

begin
  Procéder à une décomposition totale en sous-graphes appelés 'atomes' ;
  $\mathcal{K}$ est l'ensemble des séparateurs minimaux complets calculé;
  pour chaque Atome A obtenu **faire**
    si A n'est pas un graphe triangulé **alors**
      Utiliser tous les séparateurs complets de $\mathcal{K}$ intersectant C pour décomposer G;
  si Un des morceaux obtenus en décomposant G est trop gros **alors**
    CHOISIR (s'il reste des séparateurs complets) un séparateur complet central pour décomposer ce morceau.
end

5. Conclusion

L'algorithme de classification CPCL avait été intuitivement conçu pour préserver les propriétés du graphe des termes. L'implantation de cet algorithme sous la forme d'une succession de

parcours des arêtes et de réductions a permis l'obtention d'un système efficace, adapté à la classification de grands graphes creux. Mais l'intérêt d'introduire la théorie des graphes dans la classification des données textuelles ne se limite pas à la formalisation des méthodes par CAH. Nous avons illustré ici sur un exemple, comment la notion de séparateur pouvait permettre de décomposer des classes trop compactes. Nous avons aussi montré comment les récents résultats de la théorie des séparateurs minimaux permettent d'implanter très efficacement ce type de décomposition. Par la suite, nous allons tenter de valider sur un grand nombre de corpus la méthode introduite ici, qui couple classification CPCL et désarticulation des classes trop compactes. Nous espérons cependant aller bien plus loin en appliquant les algorithmes de désarticulation de graphes directement au graphe de termes, ce qui permettrait de remplacer la notion de composante connexe utilisée dans l'algorithme CPCL par celle d'articulation.

Références

Berry A., Blair J.R.S. et Heggernes P. (2002). *Maximum Cardinality Search for Computing Minimal Triangulations*. L. Kucera, Lecture Notes in Computer Science, Springer Verlag.

Berry A. et Bordat J.-P. (1997). *Decomposition by clique minimal separators*. Rapport de Recherche 97213. LIRMM.

Berry A. et Sigayret A. (2002). Representing a concept lattice by a graph. In *Proceedings Discrete Maths and Data Mining Workshop, 2nd SIAM Conf. on Data Mining (SDM'02)*, Arlington (VA), submitted to Discrete Applied Mathematics.

Jacquemin C. (2001). *Spotting and discovering terms through Natural Language Processing*. MIT Press.

Ibekwe-SanJuan F. (1997). *Recherche des Tendances Thématiques dans les Publications Scientifiques. Définition d'une Méthodologie Fondée sur la Linguistique*. Thèse de doctorat, Université Stendhal, Grenoble III.

Ibekwe-SanJuan F., SanJuan É. (2002). From term variants to research topics. *International Journal on Knowledge Organization (KO), special issue on Human Language Technology*, vol. (29/3-4).

Ibekwe-SanJuan F. et Dubois C. (2002). Can Syntactic variations highlight semantic links between domain topics ?. In *Proceedings of the 6th International Conference on Terminology and Knowledge engineering (TKE 2002)*: 57-63.

Rose D. J., Tarjan R. et Lueker G. (1976). Algorithmic aspects of vertex elimination on graphs. *SIAM Journ. Comput.*, vol. (5): 146-160.

Sander G. (1996). *Visualisierungstechniken für den Compilerbau*. Dissertation, Pirrot Verlag & Druck.

SanJuan É. et Ibekwe-SanJuan F. (2002). Terminologie et classification automatique des textes. In *Actes des JADT 2002*: 677-688.

Silberztein M. (1993). *Dictionnaires électroniques et analyse automatique des textes. Le système INTEX*. Masson.

Tarjan R. (1985). Decomposition by clique separators, *Disc. Math.*, vol. (55): 221-232.

Analyse sémantique latente et segmentation automatique des textes

Yves Bestgen

FNRS – UCL/PSOR –1348 Louvain-la-Neuve – Belgique
yves.bestgen@psp.ucl.ac.be

Abstract

Latent semantic analysis (LSA) is employed in psycholinguistics and in computational linguistics to acquire domain-independent semantic knowledge which is then used to model cognitive processes or to develop automated text analysis technologies. By analyzing the effectiveness of an algorithm of text segmentation which rests on LSA, we show not only that a generic semantic space is less effective than a specific space, but also that a series of parameters, which must be fixed at the time of the acquisition phase of the semantic knowledge, affect the performance of the algorithm.

Résumé

L'analyse sémantique latente (ASL) est employée tant en psycholinguistique qu'en linguistique computationnelle pour acquérir des connaissances sémantiques indépendantes de tout domaine qui sont ensuite employées pour modéliser des processus cognitifs ou développer des techniques d'analyse automatique du langage. En analysant l'efficacité d'un algorithme de segmentation des textes qui repose sur l'ASL, nous montrons non seulement qu'un espace sémantique générique est moins efficace qu'un espace spécifique, mais aussi qu'une série de paramètres qui doivent être fixés lors de la phase d'acquisition des connaissances sémantiques affectent les performances de l'algorithme[1].

Mots-clés : segmentation automatique, analyse sémantique latente, connaissances indépendantes de tout domaine, lemmatisation.

1. Introduction

Depuis quelques années, l'analyse sémantique latente (ASL ou LSA : Latent Semantic Analysis) est à la base d'un nombre de plus en plus important de recherches en psycholinguistique (Landauer *et al.*, 1998). Cette technique vise à construire un espace sémantique de très grande dimension à partir de l'analyse statistique de l'ensemble des cooccurrences dans un corpus de textes. Son succès en psychologie s'explique par son utilisation pour développer des simulations des processus psycholinguistiques à l'œuvre lors de la compréhension du langage (Landauer et Dumais, 1997), incluant, par exemple, un « modèle computationnel » du traitement des métaphores (Kintsch, 2000 ; Lemaire *et al.*, 2001), mais aussi l'analyse de la cohérence dans des textes (Bestgen *et al.*, 2003 ; Foltz *et al.*, 1998).

Cette technique n'est pas, tant s'en faut, l'apanage de la psychologie. Elle repose en effet sur la décomposition en valeurs singulières, une propriété des matrices rectangulaires, proposée par Eckart et Young dès 1936, qui est à la base des méthodes factorielles d'analysés de

[1] Cette recherche a bénéficié du soutient de la Communauté française de Belgique – Actions de recherche concertées et du Fonds de la Recherche fondamendale collective (FRFC/FNRS).

données (Lebart, Morineau et Piron, 2000). Plus récemment, la psychologie a emprunté cette technique à l'analyse automatique du langage (Deerwester *et al.*, 1990). Dans ce champ de recherche, l'intérêt principal de l'analyse sémantique latente est de permettre la construction automatique de connaissances sémantiques génériques (*domain-independent knowledge*). Ces connaissances peuvent donc être employées quel que soit le domaine dont sont issus les textes pour développer des techniques d'analyse du langage comme celles employées en indexation automatique de documents, en recherche de l'antécédent d'une anaphore, en identification automatique d'expressions idiomatiques ou encore en segmentation des textes (Choi *et al.*, 2001 ; Degand et Bestgen, 2003 ; Klebanov et Wiemer-Hastings, 2002). En psychologie également, la thèse selon laquelle il est possible de constituer un espace sémantique générique apte à modéliser de nombreux processus mentaux est aussi fréquemment mise en avant (Kinstch, 1998 ; Landauer et Dumais, 1997). A notre connaissance, cette thèse n'a jamais fait l'objet d'une vérification empirique. Bien plus, plusieurs recherches ont été menées en recourant à des bases spécifiques au domaine dont était issu le matériel (par exemple Bestgen et Cabiaux, 2002 ; Wolfe *et al.*, 1998). Etudier la validité de cette thèse est un des objectifs principaux de la présente étude. Nous souhaitons toutefois aller au-delà de cette seule question en étudiant l'impact d'autres paramètres qui doivent être fixés lors de la construction d'un espace sémantique et qui sont pourtant rarement discutés par les auteurs.

Pour aborder ces questions, le champ de la segmentation automatique des textes nous semble particulièrement propice pour les raisons suivantes. Tout d'abord ; l'opposition entre connaissances spécifiques à un domaine et connaissances génériques y est particulièrement prégnante (Ferret, 2002). De plus, la segmentation automatique des textes est un domaine de recherche en plein développement (Manning et Schütze, 1999 : 572) et, récemment, Choi *et al.* (2001) ont montré que l'ASL permettait le développement d'un algorithme de segmentation automatique plus efficace que plusieurs procédures classiques. Notons toutefois que ces auteurs n'ont eu recours qu'à un espace sémantique spécifique au matériel employé pour tester leur algorithme. L'analyse de la cohérence/segmentation des textes est tout aussi importante pour la psycholinguistique puisqu'identifier les relations de cohérences, mais aussi les ruptures thématiques dans un texte, est une composante centrale de la compréhension (Bestgen et Vonk, 2000 ; Gernsbacher, 1990 ; Kintsh, 1998). Enfin, les méthodes de segmentation automatique fournissent un critère relativement objectif pour jauger l'efficacité d'un espace sémantique donné, critère beaucoup plus difficile à définir dans les études psycholinguistiques.

La suite de ce rapport est structurée de la manière suivante. Après avoir décrit l'analyse sémantique latente et les paramètres qui doivent être fixés lors de la construction d'un espace sémantique, nous présentons l'algorithme de segmentation de Choi. Ensuite, nous rapportons une expérience visant à déterminer l'impact d'un ensemble de paramètres sur l'efficacité de la segmentation. Dans la conclusion, les limites de la présente étude et des pistes dé développement sont discutées.

2. L'analyse sémantique latente

Le point de départ d'une analyse sémantique latente consiste en un tableau lexical qui contient le nombre d'occurrences de chaque mot dans chacun des documents, un document pouvant être un texte, un paragraphe ou même une phrase. Pour dériver d'un tableau lexical les relations sémantiques entre les mots, la simple analyse des cooccurrences brutes se heurte à un problème majeur. Même dans un grand corpus de textes, la plus grande partie des mots sont relativement rares. Il s'ensuit que les cooccurrences le sont encore plus. Leur rareté les rend

particulièrement sensibles à des variations aléatoires (Burgess *et al.*, 1998 ; Rajman *et al.*, 1997). L'ASL résout ce problème en remplaçant le tableau de fréquences original par une approximation qui produit une sorte de lissage des associations. Pour cela, le tableau de fréquences fait l'objet d'une décomposition en valeurs singulières avant d'être recomposé à partir d'une fraction seulement de l'information qu'il contient. Les milliers de mots caractérisant les documents sont ainsi remplacés par des combinaisons linéaires ou 'dimensions sémantiques' sur lesquelles peuvent être situés les mots originaux. Contrairement à une analyse factorielle classique, les dimensions extraites sont très nombreuses (plusieurs centaines) et non interprétables. Elles peuvent toutefois être vues comme analogues aux traits sémantiques fréquemment postulés pour décrire le sens des mots (Landauer *et al.*, 1998). Les différentes étapes nécessaires pour dériver un espace sémantique d'un tableau lexical sont illustrées dans l'Annexe 1.

Tant les mots que les segments originaux sont positionnés dans cet espace sémantique, ce qui permet de mesurer leur proximité. Plus précisément, le sens de chaque mot y est représenté par un vecteur. Pour mesurer la similarité sémantique entre deux mots, on calcule le cosinus entre les vecteurs qui les représentent. Les mêmes calculs peuvent être effectués sur les vecteurs qui représentent les documents analysés. Le plus important toutefois est que cette technique est encore plus générale puisqu'elle permet de calculer le vecteur qui correspond à un groupe de mots même si ce groupe de mots ne constitue pas un document analysé en tant que tel. Il est ainsi possible d'analyser la proximité sémantique entre deux phrases que celles-ci fassent partie du corpus de départ ou non, que le corpus de départ ait été segmenté en documents correspondant à des phrases ou non.

Une série de décisions, rarement discutées dans les travaux basés sur l'ASL (mais voir Lebart et Salem (1994) pour une discussion dans le domaine de la statistique textuelle), doivent être prises dès la constitution du tableau lexical. La première porte sur la nature des documents analysés : textes entiers comme des articles de journaux ou d'encyclopédies, paragraphes, phrases, unités arbitraires de tailles constantes. A cela s'ajoute la décision de lemmatiser ou non le corpus, un facteur connu pour affecter la classification automatique des textes (Riloff, 1995), d'éliminer ou non les mots très fréquents et les mots très rares (Dumais, 1995). Selon nous, l'explication la plus probable de l'absence de discussion de ces paramètres réside dans la croyance érigée en postulat que le très grand nombre de textes et de mots qui interviennent dans ce genre d'analyse rend peu probable l'existence d'effets liés à ces paramètres. Pour vérifier le bien-fondé de cette assertion et évaluer l'impact du caractère spécifique ou non de l'espace sémantique, nous avons réalisé une expérience dans laquelle des espaces sémantiques construits en faisant varier ces paramètres ont été employés comme source de connaissance pour segmenter des textes au moyen de la méthode proposée par Choi *et al.* (2001).

3. La segmentation des textes et l'algorithme de Choi

Depuis une dizaine d'années, de nombreuses méthodes ont été proposées pour segmenter automatiquement des textes. Elles se distinguent principalement par le type d'indices employés. Certaines se basent exclusivement sur une analyse de la cohésion lexicale alors que d'autres prennent également en compte des dispositifs linguistiques qui ont pour fonction de signaler la présence de changement de thèmes. Une autre distinction importante oppose les approches qui s'appuient exclusivement sur les informations contenues dans le texte à segmenter et celles qui ont recours à des connaissances acquises par ailleurs. La méthode proposée par Choi (2000 ; Choi *et al.*, 2001) se base exclusivement sur la cohésion lexicale, mais existe dans deux versions correspondant à ce second critère de différenciation.

La procédure de Choi est composée de trois étapes. Tout d'abord, le document à segmenter est découpé en unités textuelles minimales, habituellement les phrases. Les mots composant ces phrases sont soumis à différents traitements comme la suppression de mots peu informatifs sur le thème du texte (article, pronom, verbes très fréquents, ...) et une lemmatisation. Ensuite, une mesure de similarité entre toutes les paires d'unités prises deux à deux est calculée. Enfin, le document est segmenté de façon récursive en fonction des frontières entre les unités textuelles qui maximisent la somme des similarités moyennes à l'intérieur des segments ainsi constitués.

L'étape la plus importante pour la présente étude est celle qui calcule la similarité entre toutes les paires de phrases prises deux à deux. La procédure initialement proposée par Choi (2000) reposait sur la métrique du cosinus appliquée aux vecteurs représentant les paires de phrases (Manning et Schütze, 1999 : 539 et suiv.) . Pour être déclarés cohérents, deux passages doivent contenir des mots communs. Il s'agit d'une conception très restrictive de la cohésion lexicale. Afin de dépasser cette limitation, Choi *et al.* (2001) ont proposé d'employer l'analyse sémantique latente pour estimer la similarité entre deux phrases. Pour ce faire, on applique la métrique du cosinus non aux vecteurs bruts, mais aux vecteurs pondérés par les dimensions sémantiques dérivées par l'analyse sémantique latente (Manning et Schütze, 1999 : 554 et suiv.). Les étapes ultérieures sont identiques quelle que soit la méthode employée pour calculer les similarités.

Dans une première évaluation réalisée sur un matériel très spécifique (voir ci-dessous), Choi (2000) et Choi *et al.* (2001) ont montré que leurs méthodes étaient plus efficaces que plusieurs autres approches telles que *TextTilling* de Hearst, *DotPlot* de Reynar, *Segmenter* de Kan, Klavans et McKeown et le *Maximum-probability segmentation algorithm* de Utiyama et Isahara.

4. Expérience

L'ensemble des analyses rapportées ici a été effectué sur la base d'un corpus d'articles parus dans le journal belge francophone *Le Soir* en 1997 (plus de 24 000 000 de mots). Le corpus de départ a été divisé en deux périodes : de janvier à juin (S1) et de juillet à décembre (S2). Comme les articles étaient de longueurs très variables, nous avons éliminé de chaque période les articles faisant partie des 10% les plus courts ou des 10% les plus longs, soit ceux de moins de 36 et de plus de 2219 mots dans S1 et de moins de 41 et de plus de 2252 dans S2.

4.1. *Matériel pour les tests*

Pour déterminer l'impact des paramètres décrits ci-dessus sur la précision de la segmentation, nous avons employé la méthodologie, devenue classique, avec laquelle Choi a évalué ses algorithmes (Ferret, 2002 ; Utiyama et Isahara, 2001). Elle consiste à retrouver les frontières entre des textes qui ont été concaténés. Chaque ensemble de tests est composé de 100 échantillons. Chaque échantillon est composé de 10 segments de textes. Chaque segment est composé des *n* premières phrases d'un texte sélectionné aléatoirement dans le corpus de départ. Pour tester ses algorithmes, Choi a fait varier le paramètre *n* en prenant les 3 à 5, les 6 à 8 ou les 9 à 11 premières phrases. Comme les résultats furent quasiment identiques quelle que fût la valeur du paramètre *n*, nous n'avons employé qu'un seul des trois couples : les 9 à 11 premières phrases. La valeur du paramètre *n* pour chaque segment de texte est déterminée aléatoirement. Pour chaque demi-année du Soir, deux ensembles de test, composés chacun de 100 échantillons, ont été constitués par une procédure aléatoire entièrement automatisée.

4.2. *Manipulation des paramètres pour la constitution des espaces sémantiques*

1) *Base spécifique ou non spécifique*. Les bases sémantiques étaient constituées soit à partir de S1, soit à partir de S2. Lorsque le matériel de test a été extrait des articles qui ont servi à construire la base sémantique, on parlera de condition spécifique. Dans le cas inverse, par exemple matériel de test extrait des 6 derniers mois et base construite à partir des 6 premiers mois, on parlera de condition non spécifique.
2) *Unité textuelle pour la constitution de la base*. Trois types de découpages ont été testés : textes (Texte), paragraphes[2] (Parag.) ou unités arbitraires de taille constante. Dans ce dernier cas, trois tailles arbitraires ont été analysées : unité de 50 mots (Arb. 50), unité de 250 mots (Arb. 250), unité de 375 mots (Arb. 375). La première correspond approximativement à la taille moyenne d'un paragraphe et la dernière à celle d'un texte. Les unités arbitraires respectaient les limites des textes.
3) *Lemmatisation ou non.* Nous avons employé le programme « TreeTagger » de Schmid (1994) pour lemmatiser les corpus.
4) *Suppression des mots très rares*. Trois niveaux ont été définis pour ce paramètre : absence de suppressions[3], suppression des mots dont la fréquence totale dans le corpus est inférieure à 10 ou à 20.

Un cinquième paramètre aurait pu être manipulé : la suppression ou non des mots fréquents. Il ne l'a pas été parce que les procédures de segmentation automatique des textes prévoient systématiquement la suppression de mots peu informatifs parce que très fréquents ou appartenant à des classes fermées (article, pronom, ...). Nous avons donc choisi de toujours supprimer ces mots tant dans le matériel-test que lors de la construction des espaces sémantiques. Nous n'avons pas non plus manipulé le nombre de vecteurs conservés dans l'espace sémantique réduit, le fixant à 300 une valeur classique pour ce genre d'analyse (Landauer *et al.*, 1998).

4.3. *Implémentation*

Les décompositions en valeurs singulières ont été effectuées au moyen du programme SVDPACKC (Berry, 1992 ; Berry *et al.*, 1993). L'algorithme C99 de Choi (2000) et la variante incluant l'analyse sémantique latente (Choi *et al.*, 2001) ont été implémentés en C. Il n'a pas été possible de valider spécifiquement la version ASL de l'algorithme parce que nous ne disposons pas de l'espace sémantique employé par Choi *et al.* (2001). Par contre, la version C99, qui emploie les vecteurs de mots bruts, a pu être validée en comparant les résultats obtenus avec notre implémentation à ceux rapportés par Choi (2000) pour une partie de son matériel. Une correspondance parfaite a été obtenue.

5. Analyses et résultats

Afin de simplifier les analyses, nous avons employé la variante de l'algorithme de Choi dans laquelle le nombre de segments à trouver est imposé. Pour évaluer l'exactitude d'une segmentation, nous avons utilisé les indices classiques de précision et de rappel qui, dans le cas présent, sont égaux puisqu'on impose à l'algorithme de produire le nombre correct de segments[4].

[2] Nous avons éliminé les 10% des paragraphes les plus courts et les 10% les plus longs.

[3] Plus exactement, tous les mots dont la fréquence dans le corpus est égale ou supérieure à 2 sont pris en compte, les hapax ne pouvant pas être analysés.

[4] Les analyses ont également été effectuées sur la métrique d'erreur proposée par Beefermn *et al.* (1998) et

Les analyses ont été effectuées en deux temps. Tout d'abord, nous avons étudié l'effet de la taille des documents sur l'espace sémantique en maintenant constants tous les autres paramètres sauf le caractère spécifique ou non de la base. Cette première analyse a montré que le découpage en textes donnait le meilleur résultat. Dés lors, nous avons choisi de n'effectuer la deuxième analyse que sur les découpages en textes. Lors de celle-ci, nous avons comparé dans un design factoriel complet les paramètres *Spécificité de la base*, *Lemmatisation* et *Suppression des mots rares*. La principale justification de cette procédure en deux temps réside dans le temps-calcul nécessaire pour constituer un espace sémantique.

5.1. *Analyse 1 : Impact de la taille des unités*

La première analyse porte sur l'effet de la nature et de la longueur des documents employés pour constituer l'espace sémantique. Cinq types de documents ont été comparés : subdivision en textes, en paragraphes et en unité de tailles constantes et arbitraires de 50 mots, de 250 mots et de 375 mots. Dans cette analyse, le caractère spécifique ou non de la base a également été pris en compte. Par contre, tous les autres paramètres ont été maintenus constants : absence de lemmatisation et seuil de fréquence pour les mots rares à 10. Les valeurs de précision pour les 400 échantillons de test ont été analysées au moyen d'une analyse de variance avec comme facteurs répétés la spécificité et le type de documents. Vu le grand nombre de données disponibles, un seuil de signification 10 fois plus strict que le classique 0.05 a été choisi pour toutes les analyses rapportées ici ($p<=0.005$).

	Documents				
	Texte	Parag.	Arb. 50	Arb. 250	Arb. 375
Spécifique	0.82	0.47	0.39	0.80	0.78
Non Spécifique	0.68	0.41	0.39	0.64	0.67

Tableau 1. Précision en fonction du type de document et de la spécificité de la base

On observe un effet très significatif du type d'unité et de la spécificité ainsi qu'une interaction entre ces deux facteurs. Comme le montre le tableau 1, les découpages en petites unités, qu'elles soient arbitraires ou non, donnent lieu à des performances très faibles. Il est à noter que Choi *et al.* (2001) ont obtenu, en anglais, des résultats nettement supérieurs avec des unités de cette taille. On observe peu de différences entre les trois autres unités, même si le découpage en texte donne lieu à des performances légèrement supérieures. L'impact de la spécificité de la base est également très important à l'avantage d'une base spécifique. L'interaction est largement provoquée par l'absence d'effet du facteur spécificité pour les unités arbitraires de 50 mots.

La conclusion principale de cette première analyse est qu'un découpage en grandes unités semble préférable. Il apparaît aussi que la variabilité de la taille des unités de type texte ne nuit pas à la performance. En effet, les meilleures performances sont obtenues avec ces unités dont la longueur varie fortement. Si des unités de ce type ne sont pas disponibles, par exemple lorsque le corpus est composé de textes très longs comme des romans uniquement découpés en chapitres, un découpage en unité arbitraire d'une taille suffisante produit des résultats satisfaisants. Notons enfin qu'en français aussi l'algorithme de Choi *et al.* (2001) atteint un niveau de performance très élevé.

employée par Choi, sans que les résultats ne soient modifiés d'une façon significative.

5.2. *Analyse 2*

Seul le découpage en texte a été employé pour cette analyse. Les 3 autres paramètres ont été croisés les uns avec les autres dans un design factoriel complet : *Spécificité de la base* (2) X *Lemmatisation* (2) X S*uppression des mots rares* (3), soit un total de 12 conditions correspondant à autant de bases sémantiques pour chaque période (S1 et S2). Ce même plan factoriel a été employé pour analyser les valeurs de précision obtenues pour les 400 échantillons de test au moyen d'une analyse de variance pour mesures répétées.

On obtient, comme lors de la première analyse un effet très important du paramètre *Spécificité de la base*. On observe aussi un effet significatif, mais nettement moins important, du seuil de suppression des mots rares.

Si la lemmatisation seule n'a pas d'effet, il existe une interaction significative entre ce facteur et la spécificité de la base. Comme le montre le tableau 2, la lemmatisation a un léger effet négatif lorsque les bases sont spécifiques (0.82 versus 0.80) alors qu'elle a un léger effet positif lorsqu'elles ne le sont pas (0.69 versus 0.68). Enfin, on observe une interaction significative entre la lemmatisation et la suppression des mots rares. On peut l'interpréter de la manière suivante : la suppression des mots rares a beaucoup moins d'effet lorsque le matériel a été lemmatisé. Ce résultat semble logique puisque l'effet principal de la lemmatisation est de réduire le nombre de formes rares en les regroupant sous les mêmes lemmes.

	Lemmatisation					
	Non			Oui		
	Seuil pour la suppression des mots rares					
	≥2	≥10	≥20	≥2	≥10	≥20
Spécifique	0.83	0.82	0.81	0.80	0.80	0.80
Non Spécifique	0.68	0.68	0.67	0.69	0.69	0.69

Tableau 2. Précision en fonction de la spécificité de la base, de la lemmatisation et du seuil pour la suppression des mots rares

5.3. *Analyse complémentaire*

Les deux résultats les plus importants de cette étude sont l'impact de la taille des unités et de la spécificité de la base. Si le premier s'interprète aisément, le second mérite une analyse complémentaire. En effet, deux explications différentes peuvent être données à cet effet. Rappelons d'abord qu'une même base est dans certaines analyses considérée comme spécifique et dans d'autres analyses considérée comme non spécifique selon l'ensemble de tests employé. Pour un ensemble de tests donné, les deux bases peuvent se différencier au niveau des dimensions sémantiques extraites, mais aussi au niveau des mots qu'elles contiennent. Il est donc possible que certains mots utiles pour la segmentation ne soient présents que dans la base spécifique. L'autre possibilité est que les relations sémantiques soient partiellement différentes dans les deux bases, que celles de la base spécifique soient plus adéquates pour traiter l'ensemble de test. Cette seconde explication nous semble beaucoup plus dommageable pour l'ASL puisqu'il ne suffirait pas d'augmenter le nombre de mots indexés dans une base pour la rendre efficace quelles que soient les analyses prévues.

Pour trancher entre ces alternatives, une analyse complémentaire a été effectuée. Son principe est simple : égaliser les mots employés pour mesurer la proximité entre les phrases. De cette

manière, seules les relations sémantiques dans les bases peuvent jouer. Nous avons donc effectué de nouvelles analyses dans lesquelles seuls les mots présents simultanément dans chaque paire[5] de bases spécifiques et non spécifiques sont utilisés pour le calcul des proximités. Les résultats indiquent nettement que ce ne sont pas les mots sélectionnés pour le calcul des proximités qui jouent, mais bien les relations sémantiques dans les bases. Lorsque les mêmes mots sont employés pour les espaces spécifiques et non spécifiques, la précision moyenne est de 0.81 dans la condition spécifique et de 0.68 dans la condition non spécifique, valeurs identiques à celles obtenues dans les analyses précédentes (voir tableau 2).

6. Conclusion

Avant de discuter d'une manière plus approfondie les résultats de cette étude, il est nécessaire de souligner plusieurs limites de la présente recherche qui mettent principalement en cause la généralité des résultats obtenus. Nous n'avons testé qu'une tâche (la segmentation), qu'une seule méthode (celle de Choi) dans un seul type de textes (des articles de journaux) et avec une seule méthodologie de test (la segmentation de fragments de documents arbitrairement concaténés). Même si la situation test résultant de toutes ces particularités correspond à celle employée par Choi (2000 et 2001) et par Utiyama et Isahara (2001), étudier l'impact des paramètres sur d'autres tâches et la segmentation avec d'autres types de textes et un autre matériel de test est nécessaire pour affermir l'argumentation.

Notre étude n'a pas non plus apporté de réponses à deux questions importantes. Tout d'abord, notre manipulation de la spécificité des bases sémantiques est très « légère » puisque nous avons opposé des bases construites à partir d'articles parus durant deux périodes successives d'un même journal. Cela a été suffisant pour produire des différences de performance très importantes. Jusqu'à quel niveau d'inefficacité serions-nous descendus si nous avions employé des journaux différents ou des genres de textes différents (un journal et une encyclopédie ou des œuvres littéraires) ? Des études complémentaires sont ici aussi nécessaires.

La deuxième question porte sur l'existence de différences importantes en fonction de la langue dans laquelle l'étude est menée. Comme indiqué plus haut, Choi *et al.* (2001) ont obtenu des résultats excellents avec des bases sémantiques dont les documents étaient des paragraphes. Dans notre étude, la construction d'un tableau lexical mots x paragraphes a produit des résultats médiocres. Nous avons également observé que l'algorithme C99 de Choi, qui ne s'appuie pas sur l'ASL et qui fonctionne très bien en anglais, produit en français des résultats très mauvais. Dans l'ensemble des essais effectués en variant les procédures de suppression des mots fréquents et de lemmatisation, la meilleure précision obtenue était inférieure à 0.60. Une différence d'efficacité entre l'anglais et le français a également été observée par Ferret (2002) lorsqu'il employa son système Topicoll pour segmenter des articles du Monde ou le matériel de Choi. Avant d'essayer d'expliquer ces observations par des différences inter-langues, plusieurs autres explications devraient être testées. Malgré les diverses tentatives effectuées, il est possible que les pré-traitements du matériel effectués par Choi soient plus efficaces que les nôtres. Or, ces pré-traitements affectent fortement les performances de l'algorithme tant en français qu'en anglais. L'effet inter-langues pourrait aussi résulter de différences au niveau du matériel à segmenter employé dans chacune des langues.

[5] Par paire de bases sémantiques, on entend ici deux bases identiques pour les paramètres lemmatisation et suppression des mots rares, mais extraites soit de S1, soit de S2.

En résumé, l'ensemble des paramètres manipulés dans cette étude a affecté l'efficacité de l'algorithme de segmentation. Les deux facteurs les plus importants sont le type de documents employé pour construire le tableau lexical et la spécificité de la base sémantique par rapport au matériel à segmenter. La lemmatisation et la suppression des mots rares ont aussi eu un effet, soit seul, soit en interaction. Il faut toutefois noter que l'effet de la lemmatisation est très faible, observation qui rejoint les conclusions de Lebart et Salem (1992, voir par exemple : 225-226). Dans le cas présent, elle réduit même l'efficacité de l'algorithme. Plus généralement, le fait d'analyser de très grands corpus de textes ne suffit pas à neutraliser les effets de ces paramètres.

Références

Beeferman D., Berger A. et Lafferty J. (1999). Statistical Models for Text Segmentation. *Machine Learning*, vol. (34) : 177–210.

Berry M., Do T., O'Brien G., Krishna V. et Varadhan S. (1993). *SVDPACKC: Version* 1.0 *User's Guide*. Tech. Rep. CS-93-194, University of Tennessee, Knoxville, TN, October 1993.

Berry M.W. (1992). Large scale singular value computation. *International journal of Supercomputer Application*, vol. (6) : 13-49.

Bestgen Y. et Cabiaux A.F. (2002). L'analyse sémantique latente et l'identification des métaphores. In *Actes de TALN 2002* : 331-337.

Bestgen Y., Degand L. et Spooren W. (2003). On the use of automatic techniques to determine the semantics of connectives in large newspaper corpora: an explorative study. In Lagerwerf L., Spooren W. et Degand L. (Eds), *Determination of information and tenor in texts: Multidisciplinary approaches to discourse 2003*. Nodus Publikationen : 189-202.

Bestgen Y. et Vonk W. (2000). Temporal adverbials as segmentation markers in discourse comprehension. *Journal of Memory and Language*, vol. (42) : 74-87.

Burgess C., Livesay K. et Lund K. (1998). Explorations in Context Space : Words, Sentences, Discourse. *Discourse Processes*, vol. (25) : 211-257.

Choi F. (2000). Advances in domain independent linear text segmentation. In *Proceedings of NAACL'00* : 26-33.

Choi F., Wiemer-Hastings P. et Moore J. (2001). Latent Semantic Analysis for Text Segmentation. In *Proceedings of NAACL'01* : 109-117.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K. et Harshman R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, vol. (41) : 391-407.

Degand L. et Bestgen Y. (2003). Towards automatic retrieval of idioms in French Newspaper Corpora. *Literary and Linguistic Computing,* vol. (18) : 249-259.

Dumais S.T. (1995). Latent semantic indexing (LSI): TREC-3 report. In Harman D. (Ed.), *Proceedings of The 3rd Text Retrieval Conference (TREC-3)* : 219-230.

Ferret O. (2002). Using collocations for topic segmentation and link detection. In *Proceedings of COLING 2002* : 260-266.

Foltz P.W., Kintsch W. et Landauer T.K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, vol. (25) : 285-307.

Gernsbacher M.A. (1990). *Language comprehension as structure building*. LEA.

Kintsch W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Kintsch W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin and Review*, vol. (7) : 257-266.

Kintsch W. (2001). Predication. *Cognitive Science*, vol. (25) : 173-202.

Klebanov B. et Wiemer-Hastings P. (2002). Using LSA for Pronominal Anaphora Resolution. In *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing*.

Landauer T.K. et Dumais S.T. (1997). A solution to Plato's problem : the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, vol. (104) : 211–240.

Landauer T.K., Foltz P.W. et Laham D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes*, vol. (25) : 259-284.

Lebart L. et Salem A. (1992) *Statistique textuelle.* Dunod.

Lebart L., Morineau A. et Piron M. (2000). *Statistique exploratoire multidimensionnelle* (3ième édition). Dunod.

Lemaire B., Bianco M., Sylvestre E. et Noveck I. (2001). Un modèle de compréhension de textes fondé sur l'analyse de la sémantique latente. In Paugam Moisy H., Nyckees V. et Caron-Pargue J. (Eds), *La cognition entre individu et société* (actes du colloque de l'ARCo). Hermès : 309-320.

Manning C.D. et Schütze H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Rajman M. et Besançon R. (1997). Text Mining : Natural Language Techniques and Text Mining Applications. In *Proceedings of the seventh IFIP 2.6 Working Conference on Database Semantics*. Chapam & Hall.

Riloff E. (1995). Little words can make a big difference for dext classification. In *Proceedings of the 18th Annual International ACM SIGIR. Conference on Research and Development in Information Retrieval* : 130-136.

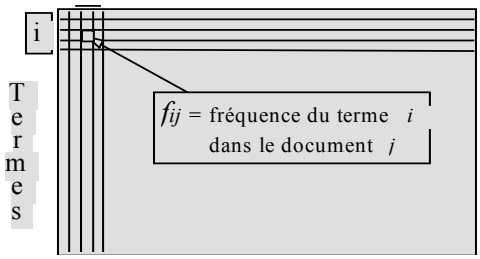
Schmidt H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. Version électronique disponible sur [http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html].

Utiyama M. et Isahara H. (2001). A Statistical Model for Domain-Independent Text Segmentation. In *Proceedings of ACL'2001* : 491-498.

Wolf M.B.W., Schreiner M.E., Rehder B. et Laham D. (1998). Learning from text: Matching readers and texts by Latent Semantic Analusis. *Discourse Processes,* vol (25) : 309-336.

Annexe 1 : Les étapes d'une analyse sémantique latente

1) Obtention d'un tableau lexical
« termes * documents »
(nombre d'occurrences de chaque
terme dans chaque document)



2) Transformation des fréquences
afin de privilégier les termes
les plus informatifs

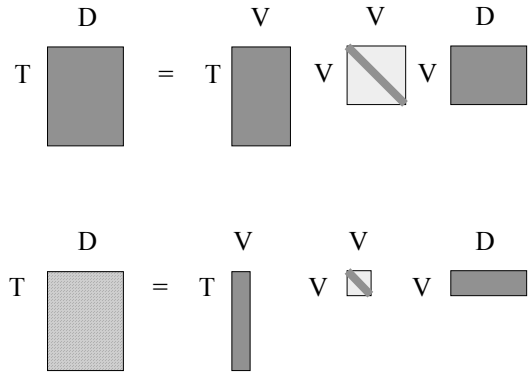
Transformation des fréquences
afin de privilégier les mots les plus informatifs

$$f_{ij}' = \frac{\log(f_{ij}+1)}{-\sum_j \frac{f_{ij}}{\sum_j f_{ij}} \log\left(\frac{f_{ij}}{\sum_j f_{ij}}\right)}$$

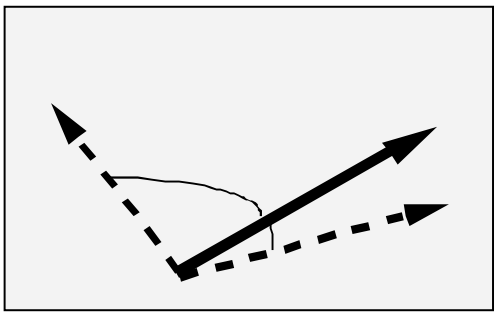
Pondération locale :
réduit l'impact des mots
très fréquents dans un document

Pondération globale :
réduit l'impact des mots « peu informatifs », c'est-à-dire qui
apparaissent avec une fréquence constante dans les documents
Exemples : mots peu informatifs : faire, pouvoir, deux, nouveau
mots très informatifs : neutrino, cancérogène, fiacre

3) Décomposition en valeurs singulières
- Compression de l'information par la sélection des k dimensions orthogonales les + importantes ($100 \leq k \leq 300$)
- Permet d'obtenir les vecteurs qui représentent les termes dans l'espace comprimé



4) Emploi :
- Calculer la proximité sémantique entre des mots ou des segments
- Le sens d'un mot est représenté par un vecteur
- La similarité entre deux mots est mesurée par le cosinus entre les vecteurs correspondants (idem pour les segments)



Un baromètre affectif effectif : corpus de référence et méthode pour déterminer la valence affective de phrases

Yves Bestgen[1], Cédrick Fairon[2], Laurent Kevers[2]

Université catholique de Louvain
[1]FNRS, Faculté de psychologie, Unité de psychologie sociale et des organisations
[2]Faculté de philosophie et lettres, Centre de traitement automatique du langage
Place Cardinal Mercier – 1348 Louvain-la-Neuve – Belgique
yves.bestgen@psp.ucl.ac.be ; fairon@tedm.ucl.ac.be ; kevers@tedm.ucl.ac.be

Abstract

The work objective of the research project presented here is to develop an Information Extraction method for selecting sentences containing named entities (person names or organisation names) and for rating automatically their affective valence (pleasant-unpleasant). In order to achieve this goal, a named entities extraction system is used together with a lexical analyser based on electronic dictionaries containing affectiveness scores. In parallel, we have built a reference corpus which permits to evaluate the scoring system by comparing its results to the judgment of human subjects

Résumé

L'objectif de la recherche rapportée ici est de développer une technique d'extraction d'information permettant de déterminer automatiquement la valence affective de phrases qui mentionnent des noms de personnalités ou de sociétés. Pour ce faire un extracteur d'entités nommées est associé à un programme d'analyse lexicale faisant appel à des dictionnaires de valence affective. Un corpus de référence est établi pour mesurer les performances du système proposé en les comparant à des jugements humains[1].

Mots-clés : valence affective, extraction d'information, entités nommées.

1. Introduction

L'objectif de la recherche rapportée ici est de développer une technique d'extraction d'information permettant de déterminer automatiquement la valence affective de phrases qui mentionnent des noms de personnalités ou de sociétés. À terme, il s'agit de pouvoir répondre à des questions comme : *le nom de tel homme politique ou de telle dirigeante d'entreprise est-il plus souvent mentionné dans un contexte linguistique plutôt positif, agréable ou bien à l'inverse dans un contexte plutôt négatif, désagréable ?* Combinée à un instrument qui extrait des journaux accessibles en ligne les passages d'articles qui mentionnent le nom d'une personne, une telle technique pourrait servir de base pour construire un « baromètre affectif ».

[1] Nous remercions pour leur aide Bernadette Dehottay et Sophie Piérard ainsi que les 10 étudiants qui ont contribué à la réalisation de cette recherche.

Elle permettrait en effet de suivre au jour le jour la valence affective des phrases dans lesquelles on fait référence à une personne donnée et d'en étudier l'évolution dans le temps.

À un niveau plus général, les questions qui portent sur la manière dont un élément est présenté et évalué dans un texte (comme de décider si la critique d'un film ou un commentaire boursier est positif ou négatif) sont particulièrement complexes pour les techniques d'extraction d'information (Wilks, 1997 ; Das et Chen, 2001 ; Pang *et al.*, 2002). Elles constituent donc un domaine de test particulièrement intéressant pour confronter différentes approches, mais aussi pour déterminer la manière la plus efficace de les combiner. Bien plus, cette complexité contraste avec la simplicité de l'approche la plus fréquemment employée pour répondre à ces questions : se baser exclusivement sur les mots qui composent un texte pour en déterminer la valence affective.

Initialement, cette approche a été développée dans le champ de l'analyse de contenu. Déjà en 1965, Heise a proposé de constituer un dictionnaire de norme d'évaluation en demandant à des juges d'évaluer sur la dimension agréable-désagréable un échantillon des mots[2] les plus fréquents d'une langue. Depuis lors, des dictionnaires pour différentes langues ont été constitués (Heise, 1965 ; Hogenraad *et al.*, 1995 ; Whissell *et al.*, 1986). Ces dictionnaires ont été employés pour évaluer la valence affective de textes, mais aussi d'unités plus petites comme des phrases (Bestgen, 1994). La procédure proposée par Heise (1965) est très simple. Dans un premier temps, on dresse la liste des mots différents et de leur fréquence dans l'unité textuelle. Cette liste est comparée à un dictionnaire qui contient un ensemble de mots dont on connaît la valence affective. À chaque fois qu'un mot se trouve dans les deux listes, on affecte la valeur indiquée dans le dictionnaire au mot du texte. Enfin, on calcule la moyenne des valeurs connues. Malgré le caractère rudimentaire de cette technique, des arguments en faveur de sa validité ont pu être apportés (Anderson *et al.*, 1982 ; Bestgen, 1994 ; Whissell *et al.*, 1986). Plus récemment, des chercheurs en extraction d'information ont adapté cette procédure afin de la simplifier en n'utilisant que des mots fortement négatifs et des mots fortement positifs (Das et Chen, 2001 ; Turney et Littman, 2002). Dans cette version, la valence affective d'un texte est déterminée en soustrayant le nombre de mots négatifs du nombre de mots positifs contenus dans ce texte.

Le caractère simpliste de cette approche basée sur les mots considérés individuellement a fait l'objet de plusieurs critiques (Bestgen, 1994 ; Pang *et al.*, 2002 ; Polanyi et Zaenen, 2003) qui plaident pour la combinaison d'informations lexicales et d'analyses linguistiques plus complexes. Par exemple, Polanyi et Zaenen (2003) soulignent la nécessité de prendre en compte les négations, mais aussi certains connecteurs (*Altough Boris is brillant in math, he is a horrific teacher*) et les opérateurs modaux (*If Mary were a terrible person, she would be mean to her dogs*). Il faut toutefois noter que les arguments empiriques à la base de ces critiques reposent sur l'analyse de quelques exemples le plus souvent construits à dessein par les chercheurs. Dépasser cette approche intuitive est nécessaire si l'on veut pouvoir évaluer les gains qu'apporte la prise en compte de chacun des dispositifs linguistiques censés affecter la valence d'une phrase.

La présente étude s'inscrit dans cette direction. Elle vise deux objectifs : tout d'abord, constituer un corpus de phrases dont la valence affective est connue. Dans ce but, nous avons sélectionné d'une manière largement aléatoire (voir ci-dessous) 702 phrases dans un corpus journalistique. Ces phrases ont été présentées à 10 juges dont la tâche était de les évaluer sur la dimension agréable-désagréable (section 3). Comme l'indique la section 3.2.1, l'accord

[2] Classes ouvertes !

inter-juges obtenu est très élevé. Ce corpus pourra donc être utilisé pour tester les différentes techniques proposées pour prédire la valence affective de phrases. Il pourra aussi être employé d'une manière plus heuristique afin de mettre en lumière les facteurs linguistiques qui modulent l'intensité affective perçue par des juges. Le second objectif de notre étude est d'évaluer sur ce corpus l'efficacité d'une approche basée exclusivement sur les mots qui composent les phrases à évaluer (section 4).

2. Constitution du corpus

1) Le matériel de départ est composé de l'ensemble des articles du quotidien national belge *Le Soir* parus entre début janvier et fin avril 1995. Une édition relativement ancienne du journal a été choisie pour éviter une trop grande sensibilité des juges par rapport à des sujets d'actualité.

2) Un système d'extraction (décrit dans Fairon et Watrin, 2002) a ensuite été utilisé pour sélectionner toutes les phrases contenant un ou plusieurs noms de personne et analyser ces séquences (c'est-à-dire reconnaître le prénom, le nom et les éventuelles informations du type profession, âge, nationalité, titre, etc.). Il s'agit d'un système basé sur des dictionnaires électroniques et des transducteurs et faisant appel au logiciel de traitement de corpus Unitex[3]. Des quelque 66500 phrases identifiées dans les textes par cette procédure, nous avons sélectionné celles d'une longueur supérieure à 8 mots et contenant au moins un nom propre apparaissant 10 fois ou plus dans l'ensemble des phrases. La justification de cette étape est que pratiquement il est peu intéressant d'étudier la valence affective associée à des personnes trop rarement mentionnées dans la presse. Au terme de cette seconde sélection, il restait un peu plus de 11000 phrases.

3) Chacune des phrases a été modifiée automatiquement de manière à remplacer le nom et la description de fonction par un prénom générique de genre adéquat (Marie, Jean, Pierre, etc.). Ceci pour éviter que les juges ne soient influencés par l'image *a priori* positive ou négative qu'ils peuvent avoir des personnes évoquées dans la presse.

4) Deux échantillons ont été extraits de ce corpus.

   a) On a extrait aléatoirement un échantillon de 700 phrases. Lors de cette étape, on a pris en compte la longueur en mots des phrases de façon a obtenir un échantillon représentatif pour cette variable. Cet échantillon est supposé représentatif du type de phrases incluant des noms propres que l'on peut trouver dans ce genre de journal. La procédure de sélection utilisée ne garantit toutefois pas qu'il contienne un nombre suffisant de phrases très agréables ou très désagréables pour pouvoir ultérieurement tester des techniques d'évaluation automatique. Pour cette raison, un deuxième échantillon de phrases a été constitué.

   b) Un deuxième échantillon de 371 phrases a été extrait du corpus et présenté à deux juges indépendants dont la tâche était d'indiquer dans quelle mesure le contenu de chaque phrase évoquait pour eux une idée plutôt désagréable, neutre ou agréable. Ils disposaient pour ce faire d'une échelle à 5 points allant de très désagréable à très agréable en passant par neutre. Sur la base de leur jugement, on a sélectionné toutes les phrases qui étaient de même polarité pour les deux juges et qu'un des deux juges au moins avait évaluées comme très désagréable ou comme très agréable.

[3] http://www-igm.univ-mlv.fr/~unitex/

5) Enfin, l'ensemble du matériel a été passé en revue par deux personnes afin d'en éliminer toutes les phrases incomplètes, syntaxiquement incorrectes ou incompréhensibles hors de leur contexte original.

Le corpus final est composé de 702 phrases dont 506 ont été sélectionnées par la procédure aléatoire et 96 par la procédure de pré-jugements.

3. Évaluation du corpus

Dix étudiants de la faculté de philosophie et lettres de l'UCL ont été recrutés, par voie d'affiche, pour participer à cette étude. Une compensation financière était offerte en échange de leur participation afin de les inciter à effectuer la tâche avec toute l'attention nécessaire.

3.1. *Modalités*

Dans les consignes, nous expliquions aux participants que nous souhaitions constituer un corpus de phrases dont l'intensité affective (agréable — désagréable) est connue. Leur tâche était de lire une série de phrases publiées dans la presse et d'indiquer si, selon eux, le contenu de chacune d'elles évoquait une idée plutôt désagréable, neutre ou agréable. Pour donner leur avis, ils disposaient de l'échelle suivante :

				neutre			
Très désagréable	**–3**	**–2**	**–1**	**0**	**+1**	**+2**	**+3**	Très agréable

Les sept échelons recevaient une dénomination verbale basée sur les mots : très (désagréable ou agréable), moyennement, un peu et neutre.

Les 702 phrases du corpus étaient présentées dans un ordre aléatoire différent pour chaque participant. La difficulté majeure de ce genre de tâche réside dans le manque initial d'ancrage de l'échelle (Bestgen, 1994). Les participants lisant les phrases une à une, il leur est difficile de situer les premières phrases par rapport aux pôles extrêmes. Afin de pallier cette difficulté, douze phrases ont été ajoutées au matériel et présentées à tous les participants dans un ordre unique au début de la tâche. Ces phrases avaient été sélectionnées dans le matériel pré-testé de sorte à exprimer l'ensemble du continuum d'évaluation depuis le pôle très désagréable jusqu'au pôle très agréable.

Concrètement, chaque participant répondait au questionnaire au moyen d'un ordinateur. Les phrases étaient successivement affichées sur l'écran juste au-dessus de l'échelle. Le participant répondait en cliquant sur le bouton correspondant à son évaluation. Un bouton « valider » lui permettait de confirmer son choix et de faire s'afficher la phrase suivante. À tout moment, il pouvait interrompre la tâche. Les consignes précisaient qu'il devait effectuer une pause d'au moins une heure vers le milieu de la tâche. En moyenne, les participants ont mis 15 secondes pour juger chaque phrase.

3.2. *Résultats*

Ce travail vise deux objectifs principaux : constituer un corpus de phrases dont l'intensité affective (agréable — désagréable) est connue et tester une technique d'estimation de la valence affective de phrases basée exclusivement sur les mots qui les composent. Ces deux objectifs ne peuvent être atteints que si un accord inter-juges suffisamment élevé est observé dans la tâche d'évaluation. Les premières analyses vérifient cette condition. Dans un deuxième temps, nous comparerons les évaluations moyennes faites par les juges aux valeurs

qui peuvent être prédites sur la base de la valence affective des mots qui composent les phrases.

3.2.1. *Accord inter-juges*

Pour déterminer l'accord inter-juges, nous avons calculé le coefficient alpha de Cronbach. Il s'élève à 0.93. Deux indices complémentaires ont été calculés. La corrélation moyenne entre les réponses d'un participant et la moyenne de tous les autres participants est de 0.75. La corrélation moyenne entre les réponses fournies par deux participants est de 0.60. Ces valeurs signalent un accord inter-juges très élevés.

3.2.2. *Analyses des évaluations*

La figure 1 présente la distribution des valeurs moyennes de valence affective pour les deux échantillons séparément. Comme le montre la distribution pour le second échantillon, la procédure de pré-jugement a bien permis d'accroître dans le corpus la proportion de phrases à forte valence. Dans l'échantillon aléatoire, on observe un large étalement des valeurs entre –2 (moyennement désagréable) et +2 (moyennement agréable).

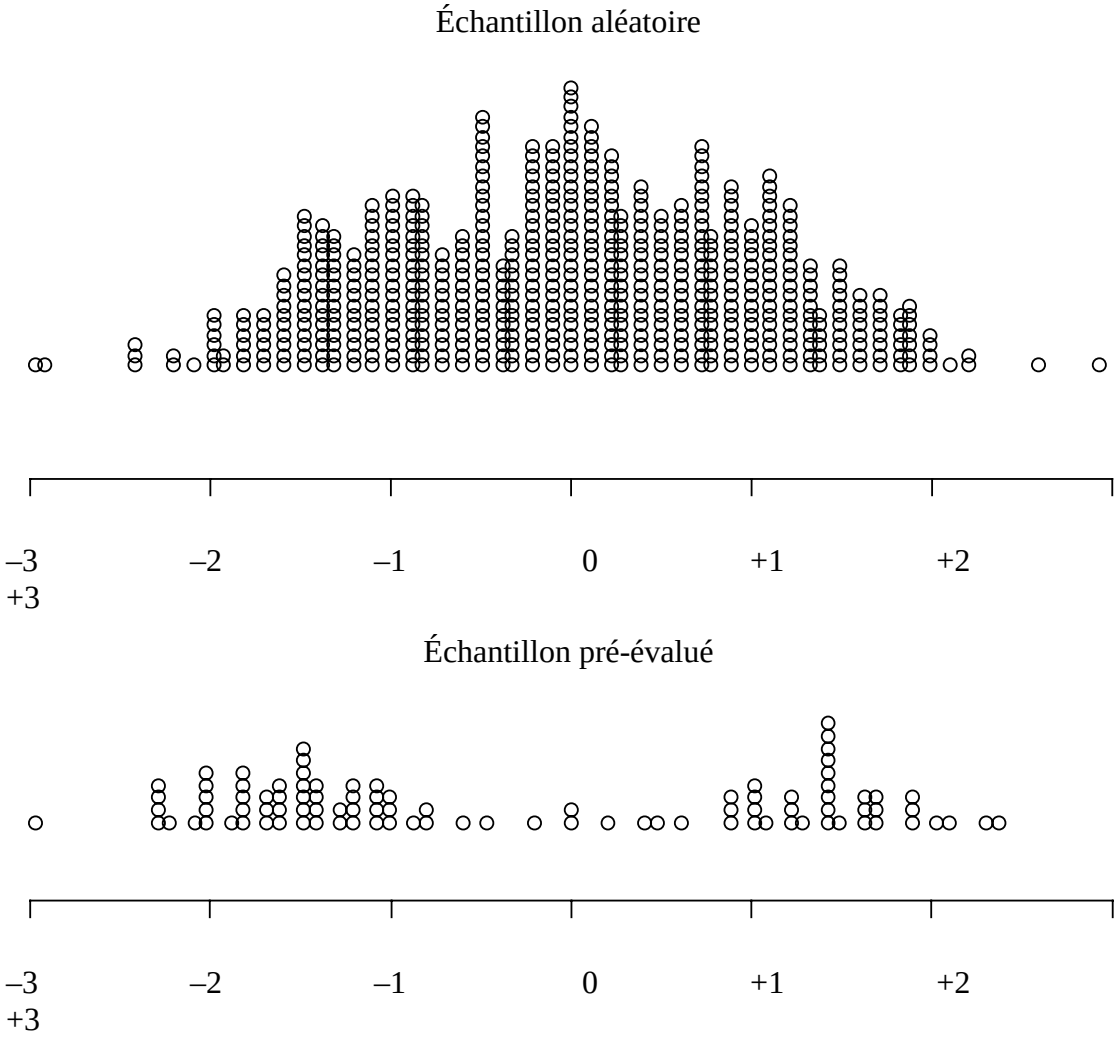


Figure 1. Distribution des valeurs moyennes de valence affective pour les deux échantillons.

4. Estimation de la valence affective

Relation entre les valeurs prédites et les valeurs réelles

Pour déterminer l'intensité affective des phrases sur la base des mots qui les composent deux dictionnaires ont été employés.

Le dictionnaire générique est composé de 3289 mots qui ont été évalués par un minimum de 15 juges sur une échelle à 7 points allant de très désagréable à très agréable (Hogenraad et Bestgen, 1989; Bestgen, 1994, 2002). Le tableau ci-dessous donne quelques exemples de mots du dictionnaire sélectionnés aléatoirement à différents niveaux d'intensité affective.

Mot	Valence	Mot	Valence
détresse	1.4	contrôlable	3.5
imbécile	1.4	outil	4.3
tristesse	1.6	risquer	4.5
hostilité	2.2	entier	4.9
impassible	2.6	revenir	5.0
superstitieux	2.8	admiratif	5.7
hâte	3.1	doux	6.0
ambigu	3.2	sincérité	6.1

Valences affectives sur une échelle allant de très désagréable (1) à très agréable (7).

La limitation principale de ce premier dictionnaire est qu'il est construit sur la base d'un échantillon de mots constitué une fois pour toutes, et donc qu'il n'est pas spécifiquement adapté au lexique employé dans le corpus analysé. Or, Bestgen (1994) a montré qu'on prédisait d'autant mieux l'intensité affective d'une unité textuelle qu'on prenait en compte l'orientation affective d'un plus grand nombre des mots qui la composent. Afin de vérifier l'impact de cette limitation, un second dictionnaire a été constitué en demandant à deux juges de parcourir la liste de tous les mots présents dans le corpus et d'en extraire ceux qui évoquent une idée agréable et ceux qui évoquent une idée désagréable. Sur la base de cette liste, l'intensité affective d'une phrase est calculée en soustrayant le nombre de mots désagréables du nombre de mots agréables contenus dans cette phrase.

Ces deux listes ont été employées pour dériver deux scores de valence affective appelés *générique* et *spécifique* selon le dictionnaire utilisé. Ces analyses ont été effectuées sur les phrases lemmatisées au moyen du programme « TreeTagger » de Schmid (1994)[4]. Seuls les mots catégorisés comme des noms, des verbes, des adjectifs ou des adverbes ont été pris en compte.

Afin d'estimer l'efficacité des deux scores lexicaux, nous les avons comparés aux valences affectives estimées par les juges au moyen du coefficient de corrélation de Pearson. Cette corrélation est de 0.39 pour le dictionnaire générique et de 0.56 pour le dictionnaire spécifique, valeurs toutes deux significatives au seuil de 0.0001. Ces corrélations ne sont pas

[4] Disponible à l'adresse http://www.ims.uni-stuttgart.de/projekte/corplex/DecisionTreeTagger.html ainsi qu'au Cental, sous forme de service en ligne (adresse http://cental.fltr.ucl.ac.be/outils.html).

modifiées lorsqu'on prend en compte la longueur des phrases ; la corrélation partielle pour le dictionnaire générique est égale à 0.37 et celle pour le dictionnaire spécifique à 0.56.

Si ces résultats sont encourageants, ils sont loin d'être parfaits. La figure 2 présente un diagramme de dispersion avec les évaluations moyennes des juges en ordonnée et les valeurs prédites par le dictionnaire spécifique en abscisse. La taille des cercles traduit le nombre de phrases qui présentent un couple de valeurs donné. Ce graphique montre qu'une partie importante de l'imprécision résulte du grand nombre de phrases qui reçoivent un score proche de zéro sur l'indice lexical. Ce constat est confirmé par les analyses suivantes. Lorsqu'on ne prend en compte que les 452 phrases dont l'indice spécifique est différent de 0 la corrélation passe à 0.62. Elle passe à 0.75 pour les 182 phrases dont l'indice spécifique est inférieur à –1 ou supérieur à 1 et à 0.83 pour les 75 phrases dont l'indice lexical est supérieur à –2 et +2. Il est à noter que cet effet s'observe tant pour l'échantillon aléatoire que pour l'échantillon pré-évalué.

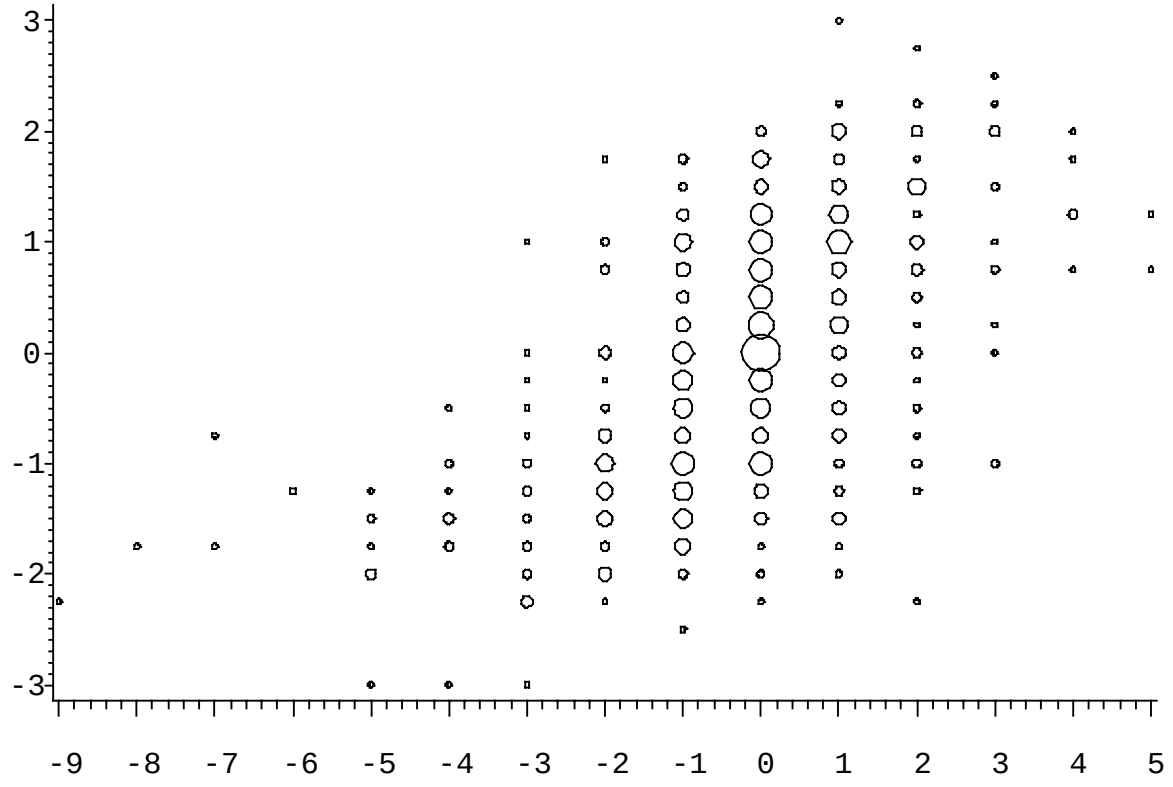


Figure 2. Diagramme de dispersion avec les évaluations moyennes des juges en ordonnée et les valeurs prédites par le dictionnaire spécifique en abscisse.

Intuitivement, on peut deviner que si une valeur très positive est obtenue par le programme pour une phrase que les juges ont évaluée très négativement, c'est que des phénomènes linguistiques échappant à la simple mesure lexicale entrent en jeu. Pour en savoir plus, nous avons passé en revue les phrases dont la valeur prédite était très différente de la valeur estimée par les juges. Nous présentons ici succinctement quelques-unes des observations que nous avons pu réaliser.

- La présence d'un verbe ou expression verbale modifie (inverse) la valeur affective de certains mots ou syntagmes : *s'en prendre à, affecter de, singer, se moquer de, hypothé-*

quer, mettre en péril, mettre en cause, etc. Par exemple, dans la phrase suivante, extraite de notre corpus, on constate la présence d'un certain nombre d'items lexicaux positifs (*fraternité, solidarité, liberté,* etc.) qui expliquent la prédiction positive donnée par le programme. Si l'avis des juges sur cette phrase est clairement négatif, c'est en raison du verbe principal de la phrase qui lui donne une valeur inverse à celle suggérée par la présence de ces mots positifs :

En frappant Tom, Rome s'en prend au-delà des clivages philosophiques, à tous les hommes et à toutes les femmes qui placent au premier rang les valeurs de fraternité, de solidarité, de liberté et de l'esprit.

– De la même manière, on peut facilement imaginer une situation opposée où un verbe inverse le caractère négatif de certains mots :

Pierre combat les inégalités et nous défend des injustices.

- Modification d'un adjectif ou d'un nom par une négation : *(moins, peu, pas, aucunement, pas le moins du monde) chanceux, heureux,* etc. ; *(Sans (aucune+la moindre)+point de+pas de) espoir, réussite, victoire,* etc.
- Ambiguïté sémantique (pouvant avoir une valeur positive ou négative en fonction des contextes) : complice (dans le crime *vs.* dans la vie).

– Nous avons également pu constater que la précision du système augmenterait s'il était capable de reconnaître :

  - des expressions/mots composés (qui très souvent n'ont pas la même valence que les éléments lexicaux dont ils sont constitués) : *centre dramatique, art dramatique, coup de foudre, etc.*
  - des noms propres à forte connotation : Front National, Brigades Rouges, etc.
  - des expressions figées ou métaphoriques comme *Au 70e de ses meetings, Jean fait un tabac à l'applaudimètre* ou *Pierre à la paupière qui papillote de joie, il est tendu d'impatience ...*

Par ailleurs, rappelons que nous souhaitons évaluer les contextes phrastiques dans le but d'établir si les noms identifiés sont évoqués dans un environnement plutôt positif ou plutôt négatif. Il est évident qu'en fonction de la place que ces noms occupent dans la structure syntaxique de la phrase, leur relation au contexte sera variable. Ceci est particulièrement remarquable dans le cas de phrases à incises : dans la phrase *« P », dit Jean*. Le nom *Jean* apparaît dans une incise que l'on peut considérer comme une parenthèse discursive. Le discours rapporté « P » et l'incise *dit Jean* appartiennent à des niveaux discursifs différents : l'incise est une intervention de l'auteur du texte qui attribue les propos rapportés à une personne. Il n'est donc pas nécessairement approprié d'appliquer les tests lexicaux sur « P » pour évaluer la situation du sujet de l'incise. En d'autre termes, une personne peut tenir des propos très marqués positivement ou négativement sans que ces propos ne la concernent directement. Il en va de même avec d'autres incises du type : *aux dires de Jean, selon Jean, d'après Jean, pour Jean, si l'on en croit Jean...*

Selon Jean, le trou du Lyonnais serait de l'ordre de 300 milliards de francs belges.

5. Conclusion

Nous avons constitué un corpus de référence permettant d'évaluer et de comparer des techniques de calcul automatique de la valence affective de phrases. Ce corpus nous a permis

de mesurer la pertinence et les limites d'une approche basée sur le lexique. Il est mis librement à la disposition des chercheurs[5] qui souhaiteraient confronter leur système d'analyse aux résultats de l'évaluation humaine.

Les résultats obtenus (corrélation pouvant aller de 0.62 à 0.83 avec les jugements humains) sont relativement satisfaisants,particulièrement dans l'évaluation d'unités textuelles aussi réduites que des phrases.

À ce stade, une limite évidente de notre système de mesures provient du fait que nous n'effectuons aucune analyse linguistique des phrases pour prendre en compte des phénomènes syntaxiques (la présence d'une négation, d'incises, etc.) ou lexicaux (mot composés à valeur négative : extrême droite) qui altère la valeur affective des phrases. Sans ambitionner une véritable « compréhension du texte » (qu'il serait d'ailleurs bien vain de chercher dans le cadre strict d'une phrase), nous pensons que l'intégration de règles linguistiques et la prise en compte d'unités lexicales composées devraient permettre d'améliorer les évaluations automatiques. Ce sera l'une des prochaines étapes de cette recherche. Nous réaliserons également l'interfaçage de notre programme avec un système de veille de la presse sur Internet[6] (Fairon 1999) dans le but de suivre l'évolution des contextes « affectifs » dans lesquels apparaissent les noms de personnalités et de proposer un baromètre politique actualisé automatiquement de jour en jour. L'implémentation d'un tel système demandera d'analyser autant que possible toutes les occurrences des noms de personnalités dans la presse et donc également de résoudre au maximum les problèmes liés à la présence d'anaphores dans les articles.

Références

Anderson C.W. et McMaster G.E. (1982). Computer assisted modeling of affective tone in written documents. *Computers and the Humanities*, vol. (16) : 1-9.

Bestgen Y. (1994). Can emotional valence in stories be determined from words ? *Cognition and Emotion*, vol. (8) : 21-36.

Bestgen, Y. (2002). Détermination de la valence affective de termes dans de grands corpus de textes. In *Actes du Colloque International sur la Fouille de Texte 2*, Nancy. INRIA : 81-94.

Das S. et Chen M.. Yahoo! for Amazon : Opinion extraction from small talk on the web, Working Paper (under review), Décenbre 2001, Santa Clara University.

Fairon C. (1999). Parsing a Web site as a corpus. In Fairon C. (Ed.), *Analyse lexicale et syntaxique : Le système INTEX*, Lingvisticae Investigationes, vol. (XXII) (Volume spécial). John Benjamins Publishing.

Fairon C. et Watrin P. (2003). From extraction to indexation. Collecting new indexation keys by means of IE techniques, EACL 2003, Budapest, Workshop on Finite State Methods.

Heise D.R. (1965). Semantic differential profiles for 1000 most frequent english words. *Psychological Monographs*, vol. (79) : 1-31.

Hogenraad R. et Bestgen Y. (1989). On the thread of discourse : Homogeneity, trends, and rhytms in texts. *Empirical Studies of the Arts*, vol. (7) : 1-22.

Hogenraad R., Bestgen Y. et Nysten J.L. (1995). Terrorist Rhetoric : Texture and Architecture. In Nissan et Schmidt (Eds), *From Information to Knowledge*, Intellect : 48-59.

[5] http://cental.fltr.ucl.ac.be/publi/2004/valenceaffective.html
[6] http://glossa.fltr.ucl.ac.be.

Pang B., Lee L. et Vaithyanathan V. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in natural language processing* : 79-86.

Polanyi L. et Zaenen A. (2003). Shifting attitudes. In Lagerwerf L., Spooren W. et Degand L. (Eds), *Determination of information and tenor in texts: Multidisciplinary approaches to discourse 2003*. Nodus Publikationen : 61-69

Schmidt H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. Version électronique disponible sur [http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html].

Turney P. et Littman M. (2002). *Unsupervised learning of semantic orientation from a hundred-billion-word corpus.* Technical Report, National Research Council Canada.

Whissell C.M., Fournier M., Pelland R., Weir D. et Makarec K. (1986). A dictionary of affect in language : IV. Reliability, validity, and applications. *Perceptual and Motor Skills*, vol. (62) : 875-888.

Wilks Y. (1997). Information Extraction as a Core Language Technology. In Paziensa M.T. (Ed.), *Information Extraction*. Springer : 1-9.

L'extraction des termes complexes : une approche modulaire semi-automatique

Ismaïl Biskri[1,2], Jean-Guy Meunier[1], Sylvain Joyal[2]

[1]LANCI – UQAM – C.P. 8888, Succ. Centre-Ville – Montréal – Québec, H3C 3P8 – Canada
[2]DMI – UQTR – C.P. 500 – Trois-Rivières – Québec, G9A 5H7 – Canada
ismail_biskri@uqtr.ca
meunier.jean-guy@uqam.ca

Abstract

Complex terms extraction systems have achieved good rates of success in the last decade. However, all these systems do not take in account users points of view, perspective, knowledge and subjectivity. Many researchers reject this fact. They argue that multiplicity of points of view leads to more than only one kind of results. In our paper we present a semi-automatic method and software tool for complex term identification. Our approach is hybrid in that it combines numeric (Bayesian approach + N-grams of words) and linguistic filters. The software tool (ESATEC) is different from other term identification tools in that it is *by design* semi-automatic: i.e. it is interactive and constantly under the user's control. The software supports the knowledge engineer's work, the (corpus) domain's expert, or the linguist, by helping them do their job more efficiently. We justify this semi-automatic approach by the need to have a more flexible and customisable tool to perform certain term identification tasks. We don't want impose on users a pseudo-standardised vision of the world. This work can be useful in terminology, indexation, information retrieval, etc.

Résumé

Durant la dernière décennie, plusieurs outils d'extractions et de repérage de termes complexes ont été mis au point. Certains de ces outils ont même été considérés comme relativement bons. Toutefois, tous ces systèmes avaient un « handicap » commun : ils ne tenaient pas compte du point de vue, de la perspective, de la connaissance et de la subjectivité de l'utilisateur. Ce que plusieurs chercheurs rejettent. Ils affirment en effet que toute opération d'interprétation ne mène pas nécessairement vers un seul type de résultat mais bien plusieurs étant donnée la multiplicité de points de vue au moment de l'interprétation. Dans notre article nous présentons une méthode ainsi qu'un outil semi-automatiques pour le repérage de termes complexes. Notre approche est hybride. Elle combine des filtres numériques (approche bayésienne + N-grams de mots) et linguistiques. L'outil mis au point, en l'occurrence ESATEC, est interactif et sous le contrôle constant de l'usager. Il assiste l'ingénieur des connaissances, l'expert du domaine ou le linguiste dans leur tâche. Nous justifions l'approche semi-automatique par un besoin d'outils flexibles et personnalisables. Nous refusons d'imposer à un usager une vision standardisée du monde. Notre travail peut être utile en terminologie, en indexation, en recherche d'information, etc.

Mots-clés : Termes complexes, approche semi-automatique, multilinguisme, apprentissage, n-grams.

1. Introduction

La langue écrite ou parlée, la traduction, le résumé, la gestion documentaire ou de l'information et bien sûr la terminologie et dans la dernière décennie l'ontologie, un repérage complet et adéquat des termes complexes dans un corpus traitant d'un domaine spécifique est considéré comme un pré-traitement des plus importants pour l'obtention de résultats d'une meilleure qualité (Strzalkowski, 1999). Dans un passé très récent, un certain nombre d'outils pour le repérage de termes ont été développés et proposés à la littérature scientifique. Ces outils acceptent comme input un texte ou corpus, généralement, pré-traité (étiqueté par exemple). Ils

produisent de façon automatique une liste de candidats termes soit au moyen d'une approche statistique (bayésienne par exemple) soit au moyen d'une approche linguistique. Les approches statistiques peuvent être multilingues. Elles sont cependant bruitées (Remaki et Meunier, 2000 ; Smadja, 1993). Les approches linguistiques sont moins bruitées, mais ne peuvent toutefois rendre compte de corpus multilingues ou certains néologismes dans des domaines spécifiques. Ces dernières approches semblent plus adaptées à des textes stéréotypés (Frath *et al.*, 2000).

La plupart des méthodes d'extraction de termes complexes préconisent l'utilisation d'un filtre linguistique pour le repérage de termes. Ce filtre utilise des patrons de termes comme ceux montrés dans Daille (1994) et Sta (1998). Dans une seconde étape, et ce pour réduire le bruit, elles utilisent des filtres statistiques ou de nature syntaxique voir sémantique comme dans Bourigault (1996) et Condamines et Rebeyrolle (2001). Les raisons qui sont données dans Daille (1994) pour justifier ce choix sont multiples :

- La perte de termes modifiés par un adverbe ou un adjectif.
- L'utilisation de filtres statistiques avant celle de filtres linguistiques induit beaucoup de bruits.
- La fréquence des termes est parfois erronée, particulièrement quand il n'y a pas au préalable une opération de lemmatisation.
- Les méthodes statistiques sont sensibles à la taille du corpus. Plus le corpus est grand plus ces méthodes donnent de meilleurs résultats.

Toutefois, malgré ces raisons, certains auteurs continuent à privilégier les méthodes statistiques, avec l'introduction de filtres linguistiques simples pour apporter des corrections aux bruits (par exemple MANTEX (Frath *et al.*, 2000)). Ces auteurs affirment que les approches linguistiques cachent pour la plupart des problèmes complexes voire majeurs.

- L'étiquetage des termes. Certains auteurs évaluent à 40 minutes le travail nécessaire pour corriger 1000 mots étiquetés.
- La lemmatisation. Le problème devant être surmonté est la nature polysémique de la langue. Le sens des mots varie très souvent en fonction du contexte dans lequel il est utilisé.
- La structure du terme. La structure syntaxique du terme est généralement considérée de type « syntagme nominal ». Pourtant cette règle souffre des exceptions et n'est pas vraie pour tous les domaines.

Le lecteur pourrait vouloir lire certains travaux relatifs au repérage des termes complexes. Nous lui conseillons, parmi d'autres, les références suivantes : Smadja (1993), Ananiadou (1994), Collier *et al.* (1998), Dagan et Church (1994), Frantzi (1997) et Lauriston (1995).

La majorité des outils disponibles dédiés à l'extraction des termes complexes ont une propriété commune : ils sont automatiques et n'interagissent que très peu avec l'usager. L'emphase mise par les programmeurs de ces outils sur leur aspect automatique cache, généralement, des pré- ou post-traitements (manuels ou pas) non-triviaux, en particulier : l'étiquetage du corpus, la lemmatisation et l'évaluation des termes candidats. Mais ce qui est particulièrement problématique est le dépouillement des résultats eux mêmes, particulièrement si on considère leur taux de rappel et de précision. La méthode et l'outil que nous présentons dans cet article permettent le repérage d'expressions récurrentes (termes) à partir d'un corpus. Toutefois, notre approche est très différente des autres :

- Par son design. Il est interactif et permet le contrôle permanent de l'usager.
- L'outil a des capacités d'apprentissage. L'identification des termes complexes est fondée sur un ensemble de termes préalablement validé par l'usager.

Le logiciel semi-automatique que nous avons développé assiste l'ingénieur des connaissances, l'expert du domaine (traité dans un corpus), ou le linguiste dans leur tâche. Dans l'esprit de nos précédents travaux (Meunier et Biskri, 2003), nous pensons que l'intervention humaine est incontournable dès lors qu'il s'agit de traitement des langues naturelles pour des résultats de haute qualité. L'identification des termes complexes n'en est qu'une modalité. La traduction ou le résumé sont d'autres exemples d'applications où le traitement automatique réalise des performances relativement pauvres comparées à des standards humains. La communauté de linguistique computationnelle ainsi que celle de l'intelligence artificielle semblent partagées en deux groupes : d'une part ceux dont l'objectif est la complète automaticité qui écarte toute intervention, d'autre part ceux dont l'objectif est d'assister intelligemment des humains dans des tâches qui ne peuvent être faites ou contrôlées que par des humains. Notre travail est définitivement représentatif du second groupe. Il y a une autre raison importante pour maintenir le contrôle humain : permettre une analyse qui tienne compte de la perspective, de la subjectivité et des connaissances du domaine de l'usager. En d'autres termes, plusieurs usagers utilisant le même outil peuvent obtenir plusieurs résultats différents : c'est ce que nous appelons une approche flexible pour le repérage des termes complexes. Les mêmes termes complexes ne sont pas nécessairement similaires par exemple en médecine et en anthropologie. C'est la raison pour laquelle la compétence d'un expert est importante. Un autre aspect que nous avons pris en considération dans notre travail a trait à l'apprentissage. Celui utilisé dans notre système est relativement simple. Il est par contre un ajout à même de favoriser le point de vue de l'usager. Notre système de repérage de termes améliore la qualité des résultats en se basant sur un ensemble de termes préalablement validés par l'usager. Cet ensemble de termes représente en soi un patrimoine qui permet d'améliorer les performances du logiciel. Enfin, le système est de conception modulaire. Chaque module (fonction) est indépendant des autres. Seul l'usager (étant donnés ses besoins) peut décider quel module exécuter. Ce genre de représentation de LATAO (Lecture et Analyse de Textes Assistées par Ordinateur) est inspiré d'un projet plus général : SATIM (Système de l'Analyse et du Traitement de l'Information Multidimensionnelle) (Meunier et Biskri, 2003).

2. Méthodologie

La méthode utilisée est hybride. Elle combine un calcul statistique bayésien avec des filtres numériques et linguistiques (nous présentons ces filtres à la section 3). La plupart de nos filtres sont « computationnellement » peu coûteux et faciles à opérer. Ils sont également facilement adaptables au traitement d'autres langues que le français et l'anglais.

Un terme complexe est considéré comme un n-gram de mots. On définit le n-gram de mots par une suite de deux mots (bi-gram), de trois mots (tri-gram) ou des fois de quatre mots (quadri-gram), voire de cinq mots (5-gram), etc. La probabilité qu'un n-gram de mots soit admis comme terme dépend de la probabilité du dernier mot de la chaîne étant donnés les mots qui le précèdent. Pour ce faire la formule générale exprimée en terme de probabilités conditionnelles, pour la reconnaissance de termes complexes est donnée dans la figure 1 ('Prob' : pour Probabilité, 'W' : pour Mot, et 'Π' : pour la multiplication).

Équation bayésienne générale :

$$\text{Prob}(W_{1\ldots n}) = \Pi_{1\ldots k}\ \text{Prob}(W_k \mid W_{1\ldots k-1})$$

Équation bayésienne pour bi-grams :

$$\text{Prob}(W_{1\ldots n}) \approx \Pi_{1\ldots k}\ \text{Prob}(W_k \mid W_{k-1})$$

Figure 1. Formules bayésiennes

Notre texte en entrée est un simple fichier texte qui n'est ni étiqueté ni lemmatisé. La seule information dont a besoin notre système est celle contenue dans les listes : liste de mots fonctionnels, liste de verbes, liste d'adverbes, etc. – ce type d'informations est dépendant de la langue mais indépendant du domaine. Le calcul bayésien détermine la probabilité des séquences ordonnées de mots dans le corpus. Les n-grams correspondent à des séquences de n mots qui peuvent correspondre à des termes complexes. En pratique, la valeur du n fréquemment utilisée est 2 (bi-grams), 3 (tri-grams), 4 (quadri-grams). Plus la probabilité d'un n-gram particulier est haute plus l'usager aura tendance à considérer ce n-gram comme terme complexe. Ainsi, la probabilité bayésienne agit comme un indicateur pour décider si un candidat terme doit être considéré comme valide ou non. Ces probabilités peuvent être perçues comme des approximations d'un phénomène linguistique complexe. Elles vont induire un taux d'erreur assez élevé, en particulier, un taux de précision assez bas. Ce qui a comme contrainte de rendre le processus de décision de l'usager plus compliqué et plus exigent en temps. Comme nous l'avons montré dans d'autres publications (Biskri *et al.*, 2003 ; Biskri et Delisle, 1999 ; Biskri et Meunier, 1998), une combinaison hybride de modèles statistiques et linguistiques peut influer positivement sur l'approche numérique pure en améliorant la granularité de leur output et ainsi leur valeur utile pour l'usager. La même idée est utilisée ici : la combinaison d'un calcul bayésien simple avec des filtres numériques et linguistiques flexibles pour l'élimination du bruit induit par le modèle bayésien de base. Un autre calcul numérique est également proposé pour permettre une phase d'apprentissage et détecter les termes déjà rencontrés et validés par l'usager.

Notre logiciel développé appelé ESATEC (pour Extraction Semi-Automatique de TErmes Complexes) est implémenté en Visual C++ et interagit avec l'usager au moyen d'une interface usager graphique. L'outil prend en entrée un corpus textuel sous format ascii, duquel il extrait le lexique. Il construit alors une matrice de collocation à partir de laquelle sont calculées les probabilités bayésiennes associées aux n-grams de mots pris dans le corpus. Dans la section suivante, nous montrons comment des filtres numériques et linguistiques sont utilisés pour améliorer la qualité des résultats.

3. Repérage des termes complexes dans ses différentes étapes

À l'étape initiale, une fois que le lexique est extrait, l'usager sélectionne les mots qui l'intéressent (il peut sélectionner l'ensemble du lexique aussi). Ces derniers peuvent être spécifiques à un domaine d'intérêt propre à l'usager. Ils forment ainsi un ensemble de mots pôles servant à déterminer uniquement les termes complexes jugés proches du domaine d'expertise de l'usager. Par exemple, dans « acide sulfurique » et « acide chlorhydrique », le mot « acide » est un mot pôle. L'usager doit aussi spécifier la taille en nombre de mots des termes complexes.

Dans la seconde étape, l'outil s'intéresse au repérage de candidat termes complexes à proprement parler en utilisant l'information donnée par l'usager. À cette étape, tout ce que nous avons est une liste de candidat termes qui sera soumise à un ensemble de filtres semi-automa-

tiques de nature numérique et/ou linguistique. Ces filtres sont indépendants les uns des autres. L'ordre dans lequel ils sont présentés n'est pas significatif. L'ordre dans lequel ils sont appliqués est par contre lui significatif. Il dépend en ce sens du choix de l'usager.

Nous donnons une brève description de ces filtres ci-après (voir également la figure 2).

Le premier filtre élimine les candidats termes qui ont une probabilité inférieure à un certain seuil donné par l'usager. Ce seuil peut faire l'objet d'une expérimentation. Il peut être possible que sa valeur soit établie après un certain nombre de tests. Le logiciel propose une valeur particulière de seuil à l'usager. Celle-ci représente la moyenne des valeurs de probabilité de l'ensemble des candidats termes. L'usager peut bien entendu prendre en considération ce seuil ou non.

Le deuxième filtre élimine les candidats termes qui commencent ou se terminent par un mot fonctionnel, un verbe, un adverbe, etc. c'est ici que l'indépendance au domaine de l'information mentionnée à la section 2 est nécessaire. L'usager peut appliquer la totalité ou juste une partie de ce filtre. Ce type de filtre a déjà été utilisé dans d'autres travaux en l'occurrence le projet Lexter (Bourigault, 1996). Il permet de déterminer les frontières d'un groupe nominal.

Le troisième filtre élimine les candidats termes qui commencent ou se terminent avec des mots spécifiques choisis par l'usager dans le lexique. Dans ce cas là, ce sont les connaissances de l'usager du domaine du corpus qui sont utiles pour éviter du bruit non productif.

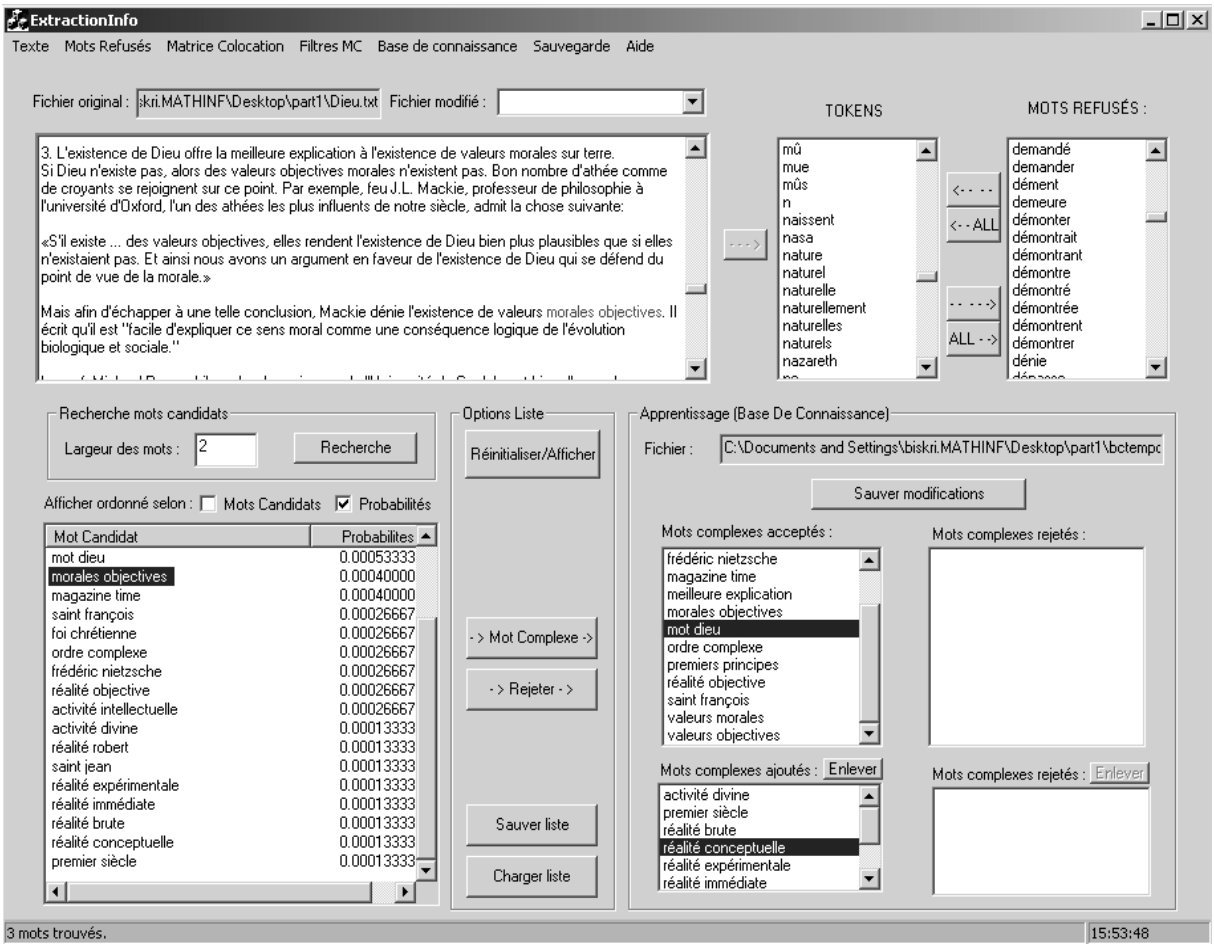


Figure 2. Interface usager et fonctionnalité

D'autres filtres peuvent être intégrés à la plate-forme, sans que cela modifie quoique ce soit au niveau de celle-ci. Certains filtres sont d'ailleurs actuellement en cours de développement. En particulier, un filtre qui consiste en l'application d'une analyse syntaxique pour éliminer l'ensemble des candidats termes qui ne sont pas des groupes nominaux (Biskri *et al.*, 2003). Des filtres similaires ont été montrés dans la littérature (Daille, 1994). Ils utilisent des patrons de termes complexes pour le français comme : Nom « de » (déterminant) Nom ; Nom « à » (déterminant) Nom ; etc. Notre analyse est basée sur le modèle de la Grammaire Catégorielle Combinatoire Applicative (Biskri et Desclés, 1997) qui a l'avantage d'être dans son essence indépendante de la langue et se prête donc, via quelques aménagements mineures à être multilingues.

4. Apprentissage

Les filtres présentés dans la section précédente nous permettent d'éliminer du bruit dans la liste des candidats termes. Ceci est en soi une « bonne nouvelle » qui permet d'améliorer le taux de précision. Toutefois, cette amélioration induit des fois une « mauvaise nouvelle » : baisse du taux de rappel. Dans le sens de cette remarque contraignante, il sera établi que le filtrage du bruit sera tributaire de la règle générale suivante : *les candidats termes ne peuvent être éliminés par un filtre s'ils ont été préalablement validés par un usager et ainsi stockés dans une table de termes complexes*. Autrement dit, les termes complexes appris par le système sont utilisés dans une ultime vérification. Aussi, plus il y a de corpus traités par le logiciel, plus la liste des termes complexes valides utilisée pour la phase d'apprentissage s'allonge et par conséquent plus la dernière vérification devient discriminante. Ce qui est intéressant avec cette dernière règle est que le taux de rappel s'en voit amélioré sans pour autant affecter le taux de précision. Cette règle peut être plus générale *: un candidat terme ne peut être éliminé par un filtre s'il dérive d'un terme complexe valide*. Par exemple, si la liste des termes valides contient « acide chlorhydrique » alors même si un candidat terme comme « acide sulfurique » apparaît avec une faible probabilité (donc susceptible d'être supprimé par le premier filtre) il ne sera pas éliminé. Une simple fonction, qui stipule que le découpage en n-grams de caractères[1] des deux termes en l'occurrence *acide sulfurique* et *acide chlorhydrique* sont similaires étant donné un certain seuil, en est garante.

Plus concrètement, pour un n = 3, le découpage en tri-grams de caractères de « acide sulfurique » et « acide chlorhydrique » donnent respectivement la suite des séquences de trois caractères suivantes : (aci, cid, ide, de , e s, su, sul, ulf, lfu, fur, uri, riq, iqu, que) et (aci, cid, ide, de , e c, ch, chl, hlo, lor, orh, rhy, hyd, ydr, dri, iqu, que). Les deux termes complexes sont alors considérés comme similaires s'ils partagent un certain nombre de tri-grams. Ce nombre devant être supérieur à un certain seuil établi par l'usager.

5. Évaluation[2]

Parce que notre approche est semi-automatique, la comparer à d'autres approches automatiques serait maladroit. Nous devons dès lors considérer notre évaluation sous un angle et une

[1] Pour rappel la notion de n-grams de caractères a été utilisée dans plusieurs travaux sur le traitement des textes oraux. Plus récemment cette notion a été au centre de l'intérêt de Greffenstette (1995) pour l'identification de la langue, et de Damashek (1995) pour le traitement du texte écrit. Ces deux chercheurs ont particulièrement prouvé que la notion de n-grams n'induisait pas de perte d'information. Des applications récentes sur l'indexation (Mayfield et MacNamee, 1998) sur l'hypertextualisation multilingue (Haleb et Lelu, 1998) et la désambiguïsation lexicale (Biskri et Delisle, 2001) confirment cela.

[2] Par souci d'homogénéiser nos évaluations mais aussi de les alléger, nous avons volontairement choisi de nous limiter aux repérages des termes complexes formés de deux mots.

perspective différents. Pour ce faire, nous avons pris un article scientifique de 20 pages, écrit en français, traitant de traitement automatique des langues.

Deux premiers résultats fondamentaux sont constatés :

- Le premier filtre qui consiste à supprimer tous les candidats termes dont la probabilité est inférieure à un certain seuil fixé (arbitrairement) par l'usager induit une baisse du taux de rappel.
- Le deuxième et troisième filtre contribuent quant à eux à améliorer le taux de précision.

Pour arriver à ce constat plusieurs évaluations sur le même texte ont été effectuées. Une première fixait un seuil de probabilité pour l'admissibilité des candidats termes de 0,001 alors qu'une seconde le fixait à 0,0005. On a pu ainsi récupérer dans le premier cas 92 candidats termes sur lesquels seuls 10 étaient acceptables, un taux de précision relatif donc de 11 %. Dans le deuxième cas, 391 candidats termes ont pu être récupérés parmi lesquels 37 étaient valables, un taux de précision relatif de 10 %. Outre le taux de précision qui est très bas dans les deux évaluations, il semble que le taux de rappel (qui reste relatif dans le cas de notre évaluation) est inversement proportionnel au seuil de probabilité. Il augmente quand le seuil de probabilité diminue. Selon nos chiffres ce taux a augmenté de 73 % en diminuant le seuil de probabilité d'acceptabilité des candidats termes de 0,001 à 0,0005.

Deux autres évaluations complémentaires aux deux premières ont été réalisées. Dans ces deux évaluations il s'agissait d'appliquer le deuxième filtre sur les listes de candidats termes obtenus après l'utilisation du premier filtre avec des probabilités d'admissibilité identiques à celles utilisées dans la première et la deuxième évaluation. Ainsi dans la troisième évaluation on applique le premier filtre avec un seuil d'acceptabilité de 0,001 puis le deuxième filtre. On obtient une liste de 10 candidats termes sur lesquels 8 sont effectivement des termes complexes. On remarque un taux de précision relatif de 80 %. Pour ce qui est de la quatrième évaluation, les mêmes deux filtres sont appliqués avec cependant un seuil d'acceptabilité de 0,0005 pour le premier filtre. Il est obtenu une liste de 41 candidats termes parmi lesquels 33 sont valides. Un taux de précision relatif de 80 % également. Il semble que le taux de précision augmente dans les deux évaluations du fait de l'utilisation du deuxième filtre. Toutefois, le taux de rappel diminue légèrement. Ainsi, dans la troisième évaluation le taux de rappel relatif diminue de 20 % alors que dans la quatrième évaluation il diminue de 10 %. Cette diminution n'est en rien dramatique. Il est possible de la corriger avec la phase d'apprentissage.

En somme ce qui ressort de ces quatre évaluations se résume en deux points :

- Le seuil de probabilité utilisé dans le premier filtre pour la validation des candidats termes doit être le plus bas possible pour garantir un haut taux de rappel.
- L'application des deuxième et troisième filtres contribuera à élever le taux de précision. Ces mêmes filtres pouvant diminuer légèrement le taux de rappel, l'utilisation de l'apprentissage permettra d'y remédier.

Ces évaluations comme on peut le constater ne tiennent pas compte encore de la phase d'apprentissage. Leur seul objectif était de montrer qu'une intrication de modules divers dans une chaîne de traitement pouvaient rendre compte de l'extraction des termes complexes et qu'on pouvait mesurer l'impact de chaque module en terme de rappel relatif et de précision relative.

Une évaluation supplémentaire plus complète s'imposait donc. Dans cette dernière évaluation nous avons soumis un texte philosophique de 20 pages à notre analyseur, tout en prenant aussi en entrée une « base de connaissances philosophiques »[3] contenant des termes complexes préalablement validés. Il en ressort que plusieurs termes complexes valides qui auraient du être éliminés par les filtres précédents (étant donnée leur nature) sont récupérés du fait de l'apprentissage du système étant donnée la base de connaissance. Aussi, le taux de rappel a pu augmenter de près de 50 %. Toutefois, le taux de précision a baissé de 7%. Cette baisse reste quand même mineure vu l'importante augmentation du taux de rappel.

Ces derniers résultats restent malgré tout bien relatifs. Ils dépendent du contenu de la base de connaissance. Or celle augmentera de taille au fur et à mesure des traitements avant de se stabiliser. Ceci aura tendance à favoriser le taux de rappel peu importe le seuil de probabilité choisi pour l'application du premier filtre. Ceci étant, notons

De cette dernière évaluation, il en ressort les termes complexes de la figure 3. Nous avons choisi les 19 premiers termes complexes par commodité pour ne pas surcharger cet article.

Meilleure explication ; Big Bang ; valeurs morales ; premiers principes ; valeurs objectives ; morales objectives ; Saint-François ; foi chrétienne ; ordre complexe ; Frédéric Nietzsche ; réalité objective ; activité intellectuelle ; activité divine ; Saint-Jean ; réalité expérimentale ; réalité immédiate ; réalité brute ; réalité conceptuelle ; premier siècle.

Figure 3. Résultats d'une évaluation sur un texte philosophique

Bien entendu, cette liste reflète nos maigres connaissances en philosophie. Une expertise en philosophie aurait certainement engendré un résultat produit au moyen de ESATEC différent. C'est d'ailleurs ce que nous voulons : que la perspective et les connaissances de l'usager influent sur le résultat.

D'autres évaluations, principalement qualitatives, que nous réservons à nos prochaines publications ont été réalisées, en particulier sur l'anglais. Des résultats presque semblable ont été obtenus. L'aspect multilingue de notre approche a pu aussi être vérifiée.

6. Conclusion

Nous avons présenté dans cet article une méthode ainsi qu'un outil semi-automatiques pour le repérage des termes complexes. Notre approche est différente de la plupart des autres outils du fait que son design est semi-automatique. Nous justifions ce choix par un besoin d'avoir un outil plus flexible et surtout plus « personnalisable ». Plus concrètement, nous voulons, dans certaines situations et dans une certaine mesure, permettre que la perspective, que les connaissances ainsi que la subjectivité de l'usager influencent le résultat. Que l'usager lise le texte pour la première fois ou pas, que plusieurs usagers considèrent le même texte, nous pensons que permettre à un même outil d'arriver à plusieurs résultats possibles représentatifs de leur état d'esprit au moment de l'analyse est d'un grand intérêt. Manifestement, cette flexibilité n'est pas possible avec la plupart des outils de repérage de termes complexes puisque la majorité produit des résultats uniques.

[3] Nous pouvons considérer autant de bases de connaissances que de domaines d'application ou d'utilisateurs. Par ailleurs nous évoquons ici une base de connaissances philosophiques. Nous précisons qu'elle a été construite pour les besoins de l'évaluation sans aucune assistance d'une expertise en philosophie.

Un autre aspect est relatif aux capacités d'apprentissages de l'outil : il peut influencer le résultat de façon significative. Comme présenté à l'évaluation, l'apprentissage peut améliorer le taux de rappel sans affecter grandement le taux de précision. En outre, la base de connaissance fait office de patrimoine que peuvent se partager les membres d'une communauté scientifique ou professionnelle quelconque.

Un dernier aspect important réside dans la méthodologie particulière qu'un utilisateur peut adopter en utilisant le logiciel. Par exemple, il pourrait préférer traiter d'abord (en partie) un corpus standard afin d'extraire une liste standard de termes complexes qui devenant disponible peut servir à l'identification suivante de termes appliquée au même corpus ou à un corpus différent. En fait, il s'avère que l'ingénieur de la langue travaille de cette manière : il établit de nouvelles connaissances sur les connaissances déjà établies.

Références

Ananiadou S. (1994). A Methodology for Automatic Term Recognition. In *Proceedings of COLING'94* : 1034-1038.

Biskri I. et Delisle S. (2001). Les n-grams de caractères pour l'aide à l'extraction de connaissances dans des bases de données textuelles multilingues. In *Actes de TALN 2001* : 93-102.

Biskri I., Meunier J.-G., Joyal S. et Gayton F. (2003). Extraction of the complex terms : the contribution of categorial grammars. In *Proceedings of PACLING'03* : 109-114.

Biskri I. et Delisle S. (1999). Un modèle hybride pour le textual data mining - un mariage de raison entre le numérique le linguistique. In *Actes de TALN 1999* : 55-64.

Biskri I. et Meunier J.-G. (1998). Vers un modèle Hybride pour le traitement de l'information lexicale dans les bases de données textuelles. In *Actes des JADT 1998.*

Biskri I. et Desclés J.-P. (1997). Applicative and Combinatory Categorial Grammar (from syntax to functional semantics). In Mitkov R et Nicolov N. (Eds), *Recent Advances in Natural Language Processing.* John Benjamins Publishing Company.

Bourigault D. (1996). Conception et exploitation d'un logiciel de termes : problèmes théoriques et méthodologiques. In *IVème journées scientifiques du réseau thématique — AUPELF-UREF* : 137-146.

Collier N., Hirakawa H. et Kumano A. (1998). Machine Translation vs Dictionary Term Translation – a Comparison for English-Japanese News Article Alignment. In *Proceedings of COLING-ACL'98* : 263-267.

Condamines A. et Rebeyrolle J. (2001). Searching for and identifying conceptual relationships via a corpus-based approach to Terminological Knowledge Base (CTKB). In Bourigault D., Jacquemin C. et L'Homme M.-C. (Eds), *Recent Advances in Computational Terminology*. John Benjamins Publishing Company.

Dagan I. et Church K. (1994). Termight : Identifying and Translating Technical Terminology. In *Proceedings of CANLP-ACL'94* : 34-40.

Daille B. (1994). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In *Proceedings of CSSALW'94* : 29-36.

Damashek M. (1995). Gauging Similarity with n-Grams : Language-Independent Categorization of Text. *Science*, vol. (267) : 843-848.

Frantzi K.T. (1997). Incorporating context information for the extraction of terms. In *Proceedings of EACL'97* : 501-503.

Frath P., Oueslati R. et Rousselot F. (2000). Identification de relations sémantiques par repérage et analyse de cooccurrences de signes linguistiques. In Charlet J., Zacklad M., Kassel G. et Bourigault D. (Eds), *Ingénierie des connaissances : Évolutions récentes et nouveaux défis.* Eyrolles.

Greffenstette G. (1995). Comparing Two Language Identification Schemes. In *Actes des JADT* 1995 : 85-96.

Halleb M. et Lelu A. (1998). Hypertextualisation automatique multilingue à partir des fréquences de n-grammes. In *Actes des JADT 1998*.

Lauriston A. (1995). Criteria for Measuring Term Recognition. In *Proceedings of EACL'95* : 17-22.

Lelu A., Halleb M. et Delprat B. (1998). Recherche d'information et cartographie dans des corpus textuels à partir des fréquences de n-grammes. In *Actes des JADT 1998.*

Mayfield J. et McNamee P. (1998). Indexing Using both n-Grams and Words. In *Proceedings of TREC7'98* : 419-424.

Meunier J.G. et Biskri I. (2003). SATIM : une plate-forme modulaire pour la construction de chaînes d'analyse de textes assistée par ordinateur. In Arnould J.C. et Blum C. (Eds), *L'édition électronique : état des lieux*. Presses de l'Université de Rouen.

Remaki L. et Meunier J.-G. (2000). Un modèle HMM pour la détection des mots composés dans un corpus textuel. In *Actes des JADT 2000.*

Smadja F. (1993). Retrieving collocations from text : Xtract. *Computational Linguistics*, vol.(19/1) : 143-178

Sta J.D. (1998). Automatic acquisition of terminological relations from a corpus for query expansion. In *Proceedings of ACM-SIGIR'98 (21st annual international conference on Research and development in information retrieval)* : 371-372.

Strzalkowski T. (1999). *Natural Language Information Retrieval*. Kluwer Academic Publishers.

Experiments on semantic categorisation of texts: analysis of positive and negative dimension

Sergio Bolasco, Francesca della Ratta-Rinaldi

Dip. Studi Geoeconomici Linguistici, Statistici, Storici per l'Analisi Regionale,
Università degli Studi di Roma "La Sapienza",
Via del Castro Laurenziano 9 – 00161 Roma – Italy
sergio.bolasco@uniroma1.it, francesca.dellaratta@uniroma1.it

Abstract

The aim of this work is the construction of a tool to categorise some properties of a text considering all evaluative words that are contained in it. For example, the tone — positive or negative — of a text can be deduced thanks to comparison with a thematic dictionary of adjectives, substantives and adverbs. The application of the dictionary to different types of languages (as dictionaries of frequency) allow us to establish a threshold point from which we can infer that a text has a negative/positive connotation.

To focus on various forms of negativity, we have carried out an analysis of all adjectives from sample of 2000 short stories by Italian students of both sexes and different age group. [1]

Keywords: text mining, categorisation, evaluation, positive and negative.

Riassunto

Scopo del lavoro è la costruzione di uno strumento di analisi che consenta di categorizzare un testo a partire dalla presenza degli elementi valutativi che vi sono contenuti. Ad esempio il tono – positivo o negativo – di un testo potrà essere dedotto grazie al confronto con un dizionario tematico di aggettivi, sostantivi e avverbi. L'applicazione del dizionario a diversi tipi di linguaggio (individuati da lessici di frequenza) consente di stabilire una soglia oltre la quale un testo può essere considerato di segno negativo. Inoltre, per focalizzare aspetti diversi della negatività si è proceduto all'analisi dell'insieme di aggettivi rilevati in un campione di 2000 di racconti prodotti da studenti italiani distinti per sesso ed età.

1. Introduction

This work is part of the field of studies tied to automatic classification of textual material by using Text Mining techniques (Sullivan, 2001). The intended experimentation will analyse the evaluative vocabulary present in some textual *corpora* to automatically characterise and classify parts of the text.

According to the usual technique of the Text Mining procedure, we will use a semantic dictionary in the Italian language as a reference model to evaluate dimension (positive *vs* negative). The comparison between the dictionary and the analysed *corpus* vocabulary will

[1] Paper realised with funds "Fondi MIUR Ateneo 2002 (C26A022374)". This work comes from a common effort; paragraphs 1 to 4 are written by Francesca della Ratta, paragraphs 5 and 6 by Sergio Bolasco and Francesca della Ratta.

allow us to identify and tag evaluative terminology contained in the *corpus*, allowing automatic classification (the positive and negative dimension) of the parts that make it up.

According to the "Pollyanna Hypothesis" formulated by Boucher and Osgood (1968), cognitive psychological studies have shown that the presence of positive terminology is much more diffused than that of a negative connotation. Text Word count finds repeatedly that positive words are used far more often than negative words, among languages and cultures as diverse as Chinese, Finnish, and Turkish (Kelly, 2003). The prevalence of positive terminology is associated with a general positive tendency, identified as a basic and universal characteristic of human nature.

Furthermore, often in the history of language, positive adjectives have had longer histories than negative ones. Also, cognitive psychological experiments have shown that it is easier to learn positive language rather than negative, both for children learning their native language and adults learning a new one (Benjafield, 1992).

2. Construction of the dictionary of evaluation adjectives

To recognise and tag evaluative adjectives in a corpus, a very comprehensive dictionary must be defined. For a reference list, we will use the English language list proposed in the dictionary of the General Inquirer (GI). It was developed in 1966 by P.J. Stone, E. Kelly and others, and recently integrated by D. Dunphy and by S. Di Cicco (Stone, 1997). It is a precious instrument for automatic classification of texts because it is based on the method of Content Analysis according to selection, and on the principle that says the frequency of certain key words is representative of the content of a text.

An instrument like the GI is very interesting for Text Mining procedures because the information contained within allows one to produce an analytical output to compile a semantic profile useful to evaluate the cognitive element and the efficacy of the text in terms of value, persuasion and emotions. The GI dictionary is composed of 13,000 lemmas of the English language, classified using different categories, that are referred both to socio-psychological theories of communication processes and to specific disciplines classifications, such as economy, law or politics. In the dictionary, the relevant role is of the "positive" and "negative" categories, that count 1,915 and 2,291 lemmas respectively; among these, adjectives alone are 590 for positive and 430 for negative.

In this study, we start with adjectives, which are considered to be the most important grammatical element to define the evaluative terminology (Marchand, 1998: 108).

To construct the dictionary, adjectives positive and negative present in GI have been translated into Italian[2] creating a list of more than 1000 lemmas. This list has been integrated with other 422 lemmas of positive and negative adjectives extracted from 6500 adjectives contained in Rep90 (corpus of 10 years of the newspaper "La Repubblica"[3]). From this list, using the dictionary present in the Taltac[4] program, all the possible adjective inflected forms

[2] The dictionary used for the translation is available in the Babylon program for translation (www.babylon.com). Throughout the translation, plurality of translation possible for each adjective has been taked into account. The translation was made at an early stage by the same person, but the list was later checked by an English mother tongue expert who is working on expanding the list of other grammar categories (nouns, verbs, adverbs).

[3] About the "La Repubblica" database see Bolasco and Canzonetti (2003).

[4] Taltac is a program for automatic lexico-textual treatment for content analysis, see Bolasco *et al*. (2000).

have been listed, which has created a dictionary of evaluative adjectives that contains about 6000 different forms.

In some cases such a list may contain some ambiguous elements whether grammatical or semantic. For instance, the word *assassino* (murderer) included in the list, may be used either as an adjective or as a noun or a verb. Similarly, the word *pure* may indicate the concept of purity or, the common conjunction synonym for also.

If the grammatical ambiguity can be overlooked, especially in view of the development of an extended version of the dictionary which will include nouns and verbs, semantic ambiguity is more complex, especially in the case of positive terminology which is often used with a neutral meaning. In the development of the dictionary it would be advisable to include terms that are potentially ambiguous, unless one can envisage how to check the list of recognised terms so as to exclude, after checking the concordance, those terms (the more frequent ones) which are incorrectly classified as positive or negative. However, in long texts, the error produced by an ambiguous classification is negligible.

3. The corpus analysed

The main corpus that was used to test the present dictionary came from the nation-wide writing competition held by the State Police in Autumn 2001 that had the theme of a story entitled "And at a certain point, the police arrived"[5] . Approximately 2,000 elementary and middle school students (between 7 and 18 years of age), from all over Italy, participated in the contest. The texts (called the Police corpus from here on) made up of all the compositions written by the children counts about one million of occurrences of words (N) and its vocabulary (V) count 47,000 words. To subject this text to the list of adjectives translated by the GI may furnish a preliminary indication of the kind of evaluative connotation in the text, and therefore it may inform us of the children's image of the police.

Besides this, it evaluates the dictionary's potential to classify texts and was be tested on different types of corpora, different writing styles, contexts and sizes to verify Boucher and Osgood's hypothesis.

4. Findings

The first step of experimentation was the comparison between the dictionary and the Police corpus' vocabulary.

The results were apparently surprising: the negative terms were more prevalent than the positive ones (with a ratio of negative to positive of 114%). This result, if the Pollyanna hypothesis is considered valid, marks a strong anomaly in the text. It is classified as having a strong negative component.

This result can probably to be attributed to noticeable structural characteristics of the text that generally start with the description of a criminal event solved by the quick intervention of the police. It can be assumed that negative adjectives are mainly associated with the events that have provoked the action of the police, whereas the positive ones have been used to describe the presence — often effective — of the police. In order to verify this hypothesis, the list of negative adjectives, taken from the text, have been analyzed.

[5] The text is being analyzed for a project Young Researchers of the University "La Sapienza" in Rome.

The most frequent negativity dimensions, or anyway those represented here above with respect to the dictionary of frequency used as reference, can be linked to the places in which the actions take place (hidden[6], dark, abandoned — *nascosto, buio, abbandonato*); to the characteristics of the "guilty person" (delinquent, murderer, dangerous, ugly, suspect, suspicious, guilty — *delinquente, assassino, pericoloso, brutto, sospetto, losco, colpevole*) and to the characteristics of the victims (poor, dead, injured, frightened, desperate, wretched, tired — *povero, morto, ferito, spaventato, disperato, misero, stanco*).

Considering the partitioned *corpus* according to categories of authors, it is possible to carry out a correspondence analysis. In our case, sex and age were chosen variables (see table 1).

The first factor is greatly determined by the comparison between the stories invented by the younger pupils (7 – 10 years old) and those invented by the older ones (14 – 18 years old). As far as the second factor is concerned, the importance of the higher age compared with the male sex of the young authors (who are also the youngest) is still decisive.

Weight		Coordinates			Absolute contributions			Squared correlations		
		1	2	3	1	2	3	1	2	3
Women	40.28	0.04	0.04	0.04	1.5	4.2	**13.7**	0.30	0.28	0.42
Men	9.72	-0.15	-0.15	-0.18	6.2	**17.4**	**56.9**	0.30	0.28	0.42
Age 7-10	13.12	-0.39	-0.07	0.08	**54.2**	5.6	14.0	**0.93**	0.03	0.04
Age 11-13	24.12	0.04	0.14	-0.06	1.0	**37.1**	13.7	0.07	**0.80**	0.14
Age 14-18	12.77	0.33	-0.18	0.03	**37.1**	**35.6**	1.7	**0.76**	0.23	0.01

Table 1. Correspondence analysis of the corpus POLIZIA

When describing the factorial planes, the most characterising adjectives in each single quadrants are examined. In the plane F1-F2, three groups of adjectives can principally be found and concentrated in the younger boys, the adolescent girls and the older boys and girls.

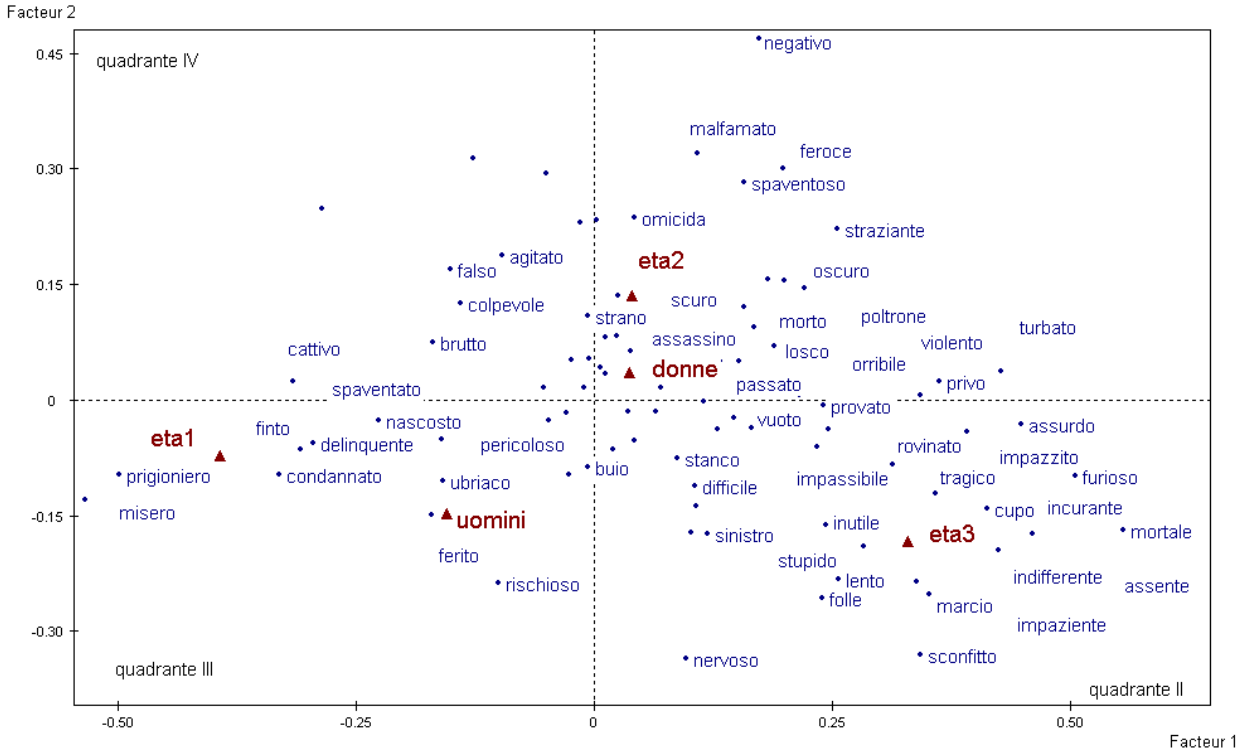
Beginning with the youngest boys, it can be noted that above all in correspondence with the quadrant III there are elementary adjectives that are used to describe the essential characteristics of the negative protagonists (bad, condemned, crafty, brusque, drunk — *cattivo, condannato, furbo, brusco, ubriaco*), the victims (prisoner, sad, injured, old, poor — *prigionero, triste, ferito, vecchio,* povero) or of the places (dark, hidden — *buio, nascosto*). These are basic adjectives, with no strong evaluative dimension, which refer to a more detached descriptive purpose from an emotional point of view. In the first quadrant, on the other hand, in correspondence with the adolescent girls, we find adjectives used for the description of the tragic situation of the victim (dead, weak, confused, upset, agonising, weary — *morto, debole, confuso, turbato, straziato, provato*) and for the elements of brutality of the negative person (terrible, murderous, assassin, frightening, evil, cruel, ferocious, negative — *terribile, omicida, assassino, spaventoso, malfamato, crudele, feroce, negativo*). In the second quadrant however, are concentrated the expressions of the older boys/girls with references to the negative protagonist, which emphasise his violent or almost demoniacal character (crazy,

[6] For clarity of description, the reference to the lemmas is given here.

mad, hateful, sinister, rotten, dirty, violent, threatening — *folle, impazzito, maledetto, sinistro, marcio, sporco, violento, minaccioso*).

The factorial analysis offers an additional key to the interpretation of the three thematic categories of adjectives (negative characters, victims, places in which the event occours), which can be linked back to the emotional dimension and varies from the description of the victim's experience to the emphasis of the "demoniacal" effects of the negative characters.

It is of particular interest to see that the youngest children are the ones who use poorer language from an emotional point of view: perhaps because the stereotypes of a fictional origin in positive and negative events are stronger in this group of children. The emotional detachment could also be a sign of the growing and often a-critical exposure to episodes of violence among the younger children, which trigger off the uncertainty of the limits between fiction and reality.



Graph 1.+ Factorial plane of negative adjectives of the Police corpus

In a similar way it is interesting to analyze the positive adjectives. These are mainly referred to **positive protagonists**, both the police forces (the policemen-*poliziotti*, but also the detective, the investigator-*investigatore*, the inspector-*ispettore*, the guard-*guardia*), and the co-protagonists, that is to say boys and girls or other characters (neighbours, teachers, passers-by, citizens — *vicini, insegnanti, passanti, cittadini*) who are given the role of protagonist in the short stories. Among the positive adjectives quoted above, those referred more often to the positive protagonists are <good, brave, strong, kind, quick in action, heroic> *(bravo, coraggioso, forte, gentile, tempestivo, eroico*). Then there are adjectives referred to **situations** which represent either the happy ending of the story or the situation at the beginning of the story that is interrupted by the criminal action (happy, tranquil, calm, quiet, carefree — *felice, tranquillo, calmo, quieto, s*pensierato) and those referred to the **objects** stolen (precious,

priceless — *prezioso, inestimabile*)[2].

5. Comparison between different corpus

To fully evaluate the particularity of the results obtained, the second step of this study regards applying the dictionary to other corpora[7] types, which will show that, regardless of the type of text, positive adjectives are always more common than negative, confirming what has been affirmed by Boucher and Osgood. As it is shown in the table 2 below, the negativity index (Occ. Neg/Pos*100), in another analyzed corpora increases from 8% (in the case of the information given by the school to explain the education plan to students) to a maximum of high of 58% in the case of citizen's complaints on city services, another text with strong negative connotation. This result allows us to affirm that the dictionary is a useful instrument to characterize the negative level of a text.

CORPUS	N	V	Neg/Pos*100	Negative		Positive	
				n	**0/00 on N**	n	**0/00 on N**
1 – Police	989.785	47.354	**114,34**	13.804	13,95	12.073	12,20
2. Complaints taken from the Internet	75.361	12.587	**58,24**	866	11,49	1.487	19,73
3. Press Reviews on the World Cup 90	250.463	22.717	**52,04**	2.139	8,54	4.110	16,41
4. Focus group - temporary workers	20.691	3.460	**35,93**	83	4,01	231	11,16
5. Open questions- graduates on their thesis difficulties	48.730	4.529	**32,74**	184	3,78	562	11,53
6 Interviews with university student parents	63.548	5.545	**21,75**	244	3,84	1.122	17,66
7. Focus group teachers on nursery school	131.254	8.806	**21,44**	378	2,88	1.763	13,43
8. Open questions- graduates on satisfaction and dissatisfaction	52.684	4.702	**19,78**	180	3,42	910	17,27
9. Focus group IRRE teachers	80.558	8.154	**19,42**	209	2,59	1.076	13,36
10. Documents POF Lazio schools	92.432	10.013	**8,84**	135	1,46	1.527	16,52

Table 2. Comparison result on evaluative dictionary, ordered per negativity index

[2] The analysis of positive adjectives however presents greater problems of semantic and grammatical ambiguity, due to the natural prevalence of positive terminology which more often gives these terms a neutral meaning.

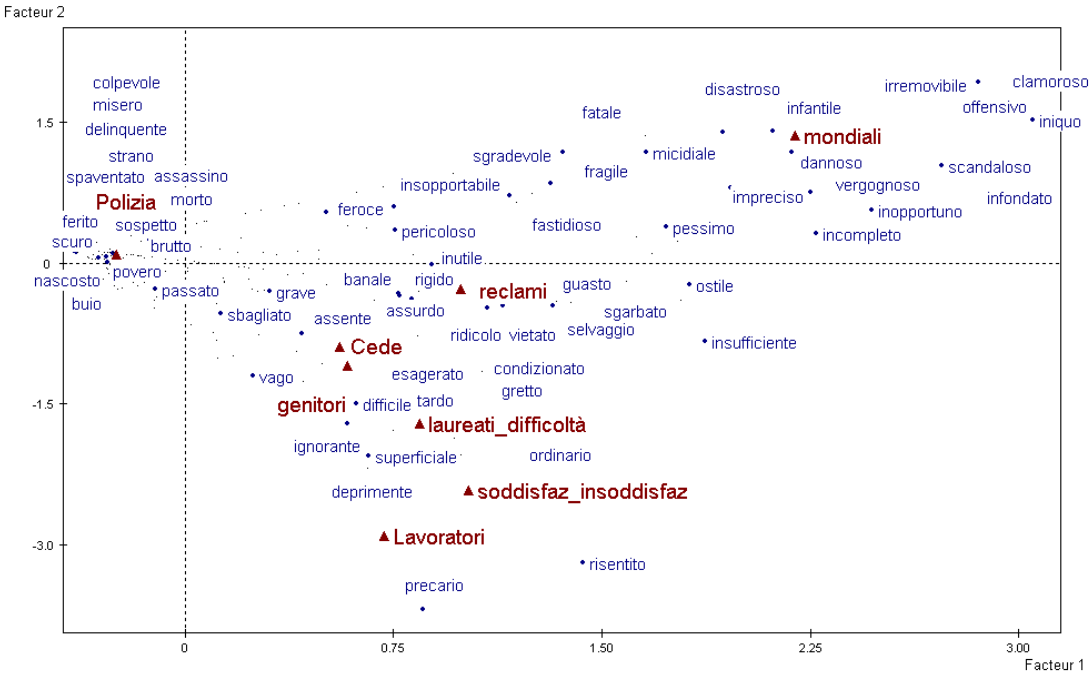
[7] The analyzed *corpora* have been collected from the writers or kindly given by other researchers. They refer to: 2) complaints taken from the Internet gathered by F. Cassoni, a university student; 3) press reviews on the World Cup '90 written by graduate students FSeM; 4) registrations of 3 focus groups on temporary workers done by Dr. L. Spera; 5) answers to open questions given to about 400 Sociology graduates in Rome in the academic years of 1997/98 on the difficulties encountered doing their thesis; 6) 30 interviews of university student parents on the future expectations for their sons; 7) 30 focus groups conducted on teachers about the quality of their nursery school; 8) the answers to open questions given to Sociology graduates on their reasons for satisfaction and dissatisfaction in their acquired training; 9) 6 focus groups on teachers on their evaluation of the nutrition education plan from the IRRE of the Lazio region; 10) documents regarding the "Plans of Educational Offer" (POF) from 16 schools in the Lazio region (10).

Furthermore, if the division of negative adjectives in the different texts undergoes a correspondence analysis, it is possible to represent the negativity dimension in some of its facets.

It is particularly interesting to note that the first factor is characterised by the opposition between the Police corpus and those of the Press Reviews on the 1990 World Cup, together with the Complaints. Such an opposition can also be brought back to the age variable, referred to the authors of the texts (the young students vs the more adult authors of the press articles). This is probably a question of the more emphasised negativity dimensions present in the texts used for the comparison, placing one in the **tragic narrative component** of the short stories on the Police (with their contents of fear, desperation, ugliness and violence) and the other in the **disgust and protest** component characterising the Press Review on the World Cup and the Complaints, with adjectives referring to scandal, negligence and rudeness, ordinariness and mediocrity, awkwardness and inappropriateness.

The second factor, on the other hand, appears to be characterised by the type of texts, with the classical comparison between written and spoken language. In fact, if the positive semi-axis is characterised by the **impersonal language** of the press, the negative one is determined by more **personal** references typical of the reviews of the focus groups, open questions or interviews. These adjectives can be interpreted as the comparison between objectivity and subjectivity: if, with regard to objectivity, there are references to harmfulness, incompetence, childishness, inaccuracy, nuisance and obscenity, with regard to subjectivity there are references to anxiety, absurdity, discontent, ordinariness, difficulty, precariousness, insecurity and embarrassment.

Of course these categories of negativity depend on the type of texts that have been analysed. In this case, it seemed interesting to demonstrate that by using the dictionary it is possible not only to order different texts according to their negativity, but also to illustrate the different negative components characterising them. If used on texts that are coherent in terms of context, such as for example political speeches or newspaper articles, this comparison could give even more interesting results with regard to the discrimination of objects.



Graph 2. Factorial plane of negative adjectives in all the corpora

After the analysis of negative dimension, the problem that has to be solved is how to define the threshold point from which on we can say a text has a negative connotation.

A possible solution could be to apply the positive-negative dictionary to some frequency dictionary in the Italian language. In this case as well, there is confirmation of the Pollyanna Hypothesis, with a negative index that varies from 50% of the enormous dictionary of "La Repubblica" newspaper to 40% of the POLIF dictionary, inserted as a reference model in the TALTAC program. This result allows us to affirm that texts with negative index higher than 40% can be considered to have a decisive negative connotation.

However index values slightly lower than 40% must be evaluated cautiously since they contain negative elements.

6. Conclusions and future perspectives

The instrument that has been defined, even though limited to the analysis of adjectives only, makes it possible to supply significant information on texts in relation to their degree of negativity. The analysis of the adjectives identified and the comparison between negative adjectives contained in different corpora then allows a more careful examination of the semantic connotation with the negative dimension. Over the coming year the dictionary will be completed with the introduction of other terms, substantives and adverbs, which will make it more complete, resolving a great part of the problems of grammatical ambiguity.

Finally, from the point of view of the automatic categorization, the possibility to tag the positive and negative adjectives in the text can permit the automatic classification of parts of the text on the basis of the greater or smaller concentration of negative terms, making it easier to extrapolate parts that present greater characteristics of negativity.

References

Benjafield J.G. (1992). *Cognition*, Prentice Hall Inc. It trad. (1995), *Psicologia dei processi cognitivi*. Il Mulino.

Bevilacqua E., della Ratta-Rinaldi F. and Orsini A. (2003). *'Quando ad un tratto arrivò la polizia.' Un viaggio nell'immaginario giovanile*. Progetto Giovani Ricercatori de "La Sapienza", dipartimento DISC.

Bolasco S., Baiocchi F. and Morrone A. (2000). *Taltac. Trattamento automatico lessico-testuale per l'analisi del contenuto*. Cisu.

Bolasco S. and Canzonetti A. (2003). *Some insight on the evolution of 1990s' standard Italian, by Text Mining techniques and automatic categorization using the lexicon of the daily "La Repubblica"*. CLADAG.

Boucher T. and Osgood C.E. (1969). The Polyanna Hypotesis. *Journal of Verbal Learning and Verbal Behavior*, vol. (8): 1-8.

Hatzivassiloglou V. and McKeown K. (1993). *A Quantitative Evaluation of Linguistic Test for the Automatic Prediction of Semantic Markedness*. Columbia University. http://acl.ldc.upenn.edu/P/P95/P95-1027.pdf.

Kelly M.H. (2003). *Naming on the bright side of life*. University of Pennsylvania, http://www.sas.upenn.edu/~kellym/brightSide.html.

Kerbrat-Orecchioni C. (1981). *L'énonciation de la subjectivité dans le langage*. Armand Colin.

Krippendorff K. (1980). *Content Analysis. An Introduction to its Methodology*. Sage. It trad. (1983), *Analisi del contenuto. Introduzione metodologica*. ERI.

Marchand P. (1998). *L'Analyse du Discours Assistée par Ordinateur. Concepts, Méthodes, Outils*. Colin.

Stone P.J. (1997). Thematic text analysis: new agendas for analyzing text content. In Roberts C. (Ed.), *Text Analysis for the Social Sciences*, Lawrence Erlbaum Associates.

Sullivan D. (2001). *Document Warehousing and Text Mining. Techniques for Improving Business Operations, Marketing and Sales*. Wiley.

Les apports de l'analyse textuelle pour l'analyse électorale : les questions ouvertes du panel électoral de 2002

Mathieu Brugidou[1], Nadine Mandran[2], Michel Moine[3], Annie-Claude Salomon[2]

[1]CIDSP, GRETS-EDF – 92141 Clamart Cedex – France
[2]CIDSP, IEP – BP 48 – 38040 Grenoble Cedex 09 – France
[3]LabSAD-UPMF, CIDSP – BP 47- 38040 Grenoble Cedex 09 – France
mandran@cidsp.upmf-grenoble.fr, michele.moine@iutz.upmf-grnoble.fr

Abstract

During the survey « Panel électoral français 2002 », the answers to several open ended questions were collected. The analysis of the texts of the responses to questions concerning the reasons of the defeat of the left, and the reasons of the victory of the right are analyzed in this document. Two methods will be used to analyze the structure of this text. A lexical analysis offer the opportunity to extract the different speeches. This step will be improve by a thematic analysis. This methodology allows a sociological and political analysis too. In fact, each interviewed person will be characterized by the different thematics of his response.

Résumé

Lors de la dernière vague de l'enquête électorale du « Panel électoral français 2002 », les réponses à plusieurs questions ouvertes ont été recueillies. Les données analysées sont les réponses aux deux items « Selon vous, pourquoi la gauche a-t-elle perdu ? » , et « selon vous, pourquoi la droite a-t-elle gagné ? ». Deux méthodes complémentaires seront utilisées pour mettre à jour la structure ce corpus. Une analyse lexicale extrait les principaux discours présents dans l'opinion. Cette étape est enrichie par une analyse thématique qui affine les classes d'énoncés. Elle permet une analyse individuelle des thèmes et une caractérisation socio-politique de ces discours.

Mots-clés : question ouverte, analyse du discours, analyse lexicale , analyse thématique, analyse électorale.

1. Présentation du Panel électoral français 2002

Les données du « **Panel électoral français 2002** » (PEF, 2002) ont été produites par le CEVIPOF, le CIDSP, le CECOP avec le soutien du Ministère de l'Intérieur et de la FNSP. Ce dispositif d'enquête a permis de mesurer les opinions et les comportements politiques des français. Cette étude a été menée en trois temps : une première vague pré-présidentielle, une seconde post-présidentielle et une troisième post-législative. Trois questionnaires comportaient principalement des questions fermées mais aussi différentes questions ouvertes. Dans la vague 1, elles portaient sur les raisons de la participation ou non au vote et sur les raisons du choix du candidat. Dans la vague 2, les questions ouvertes abordaient les problèmes d'environnement d'une part et d'autre part les raisons du vote J. Chirac, du vote J.-M. Le Pen et de l'abstention. En dernière vague, ce sont les raisons de l'abstention, de la défaite de la gauche ou de la victoire de la droite qui ont été recueillies.

2. Objectif de l'étude

Lors de la troisième vague, l'une des deux questions ouvertes suivantes était posée aléatoirement à chaque enquêté[1] :

Selon vous, pourquoi la gauche a-t-elle perdu ?
Et, selon vous, pourquoi la droite a-t-elle gagné ?

À partir de ces deux questions, un double objectif a été défini : le premier est d'isoler les différents discours produits par les enquêtés interrogés à propos de la victoire de la droite ou la défaite de la gauche à l'issue de la séquence électorale de 2002, le second est de caractériser le répondant par l'ensemble des thèmes qu'il a évoqué. La première étape consiste à analyser les corpus de chacune des questions, puis à dégager les champs lexicaux et leurs associations. Dans une deuxième étape, une liste de thèmes est construite à partir de ces champs lexicaux. La comparaison des distributions des thèmes abordés par les enquêtés permet de mettre en évidence des différences attendues selon la formulation de la question (victoire ou défaite). Mais au-delà de ce constat, les mêmes thématiques sont présentes dans les deux corpus, la différence se fait sur leur fréquence d'apparition. Enfin, les croisements des thèmes abordés par les locuteurs, avec d'une part leurs caractéristiques sociodémographiques et d'autre part leurs opinions et leurs comportements politiques ont conduit à la conclusion que certaines relations étaient indépendantes de la question posée.

3. Analyse textuelle, définition des champs lexicaux

3.1. *Analyse de discours des réponses à la question « Pourquoi la gauche a-t-elle perdu ? »*

L'analyse statistique produite par le logiciel Alceste conduit à un classement des énoncés relatifs à la défaite de la gauche, une réponse étant découpée en séquence d'énoncés.

La classification descendante fait émerger deux pôles majeurs pour expliquer la défaite de la gauche : d'une part, un constat négatif sur les actions du gouvernement et des promesses non tenues, d'autre part les motifs énoncés font référence à la perception de la campagne du premier tour de la présidentielle et aux comportements électoraux lors de cette élection. Cette analyse distingue donc quatre classes de discours.

3.1.1. *Une absence de programme et de leader (classe 1)*

L'opinion selon laquelle la raison de la défaite de la gauche est une absence de programme et de leader s'exprime dans cette classe. Cette absence est d'autant plus forte que les candidatures à gauche lors de la présidentielle étaient nombreuses, ce qui a brouillé les signaux lors de la campagne électorale. « *car elle était mauvaise pas de programmes clairs et pas de dynamique* » . « *manque d'union manque d'information et par manque de leader digne de ce nom pas de figure charismatique* ».

3.1.2. *Un vote contestataire et l'abstention (classe 2)*

Les raisons de la défaite de la gauche dans cette classe sont doubles. Nous trouvons d'une part un discours qui met en avant le vote contestataire car les individus ont été déçus des actions

[1] Le nombre de réponses à cette question ouverte est de 1003 pour la première question, de 1005 pour la seconde. Le taux de non-réponse à ces deux questions est très faible, puisque le nombre total d'enquêtés de la troisième vague est de 2013. Avant de procéder à cette étude, nous avons vérifié que la structure par âge, sexe, niveau d'études, vote au premier tour est indépendante du mode de questionnement. Les questions ouvertes ont été recueillies et retranscrites directement par l'enquêteur lors du terrain téléphonique.

menées par la gauche : « *il y a beaucoup de déçus qui ont voté sanction au premier tour et ça fait un vote d'extrême droite* ». D'autre part, c'est l'abstention qui est évoquée pour expliquer les raisons de la défaite de la gauche, « *parce que il y a eu beaucoup trop d' abstentions. Je vois surtout cela, les gens ont dit qu' ils n' iraient pas voter, même sans faire de politique, on peut avoir son opinion* ». Aussi, une des caractéristiques de cette classe est la présence de l'expression « extrême droite ». Les propos tenus ici sont une traduction des événements du 21 avril 2002 : un vote contestataire et une abstention très forte ont conduit au rejet de L. Jospin et à la victoire de l'extrême droite.

3.1.3. *Des promesses non tenues (classe 3)*

La raison de la défaite de la gauche exposée dans cette classe de discours est le fait que le gouvernement n'a pas tenu ses promesses. Ce discours fait référence à un passé lointain qui prend ancrage lors de la campagne des législatives de 1997. Cette classe est caractérisée par des marqueurs de personnalisation comme « leurs » ou « ses », qui font référence aux personnalités en place. « ils ont menti pendant trop longtemps, ils nous ont berne ils ont pas tenu leurs promesses ».

3.1.4. *Le bilan négatif des actions de la gauche. (classe 4)*

La caractéristique de ce discours est de lister et de dresser un bilan négatif des actions menées par le gouvernement. Les trente cinq heures arrivent en tête pour expliquer la défaite de la gauche. « *Les Français étaient pas satisfaits et notamment les trente cinq heures et la baisse du pouvoir d' achat, ils n'ont pas les moyens de profiter de la RTT* ». Ces déclarations mettent aussi en exergue les différents problèmes qui n'ont pas été abordés et/ou résolus par le gouvernement (retraite, sécurité, chômage,…). Ces actions sont parfois qualifiées de « *bêtise* », ou d'« *erreur* ».

3.2. *Analyse de discours des réponses à la question « Pourquoi la droite a-t-elle gagné ? »*

L'analyse statistique produite par le programme Alceste fait ressortir quatre classes de discours dans le corpus des réponses relatives à la victoire de la droite. Ces quatre classes s'organisent selon trois pôles : la peur de l'Extrême Droite, le refus de la cohabitation (classes 1 et 2), le ras le bol qui pousse au changement (classe 3) et la déception vis-à- vis de la gauche : bilan factuel des actions de la gauche et jugement critique de son action (classe 4).

3.2.1. *Le refus de la cohabitation, une meilleure campagne à droite (classe 1)*

Le contenu de ces énoncés renvoie à la nécessité de « *donner une majorité parlementaire au Président de la République afin d'éviter une nouvelle cohabitation* ». La victoire de la droite est présentée comme étant « *la suite logique de l'élection présidentielle* ». La campagne de la droite en 2002 est jugée meilleure – « *Ils avaient les meilleures têtes, le meilleur programme* », « *la droite était unie et non divisée* » – que celle de la gauche – « *décevante* », « *divisée* », « *mauvaise campagne* »,…

3.2.2. *La peur de l'extrême droite (classe 2)*

Le thème de la peur de l'extrême droite est très présent dans cette classe. La surprise devant les résultats du premier tour des présidentielles est évoquée. Le rôle de l'abstention au premier tour de la présidentielle, des reports des votes « de gauche » au deuxième tour de la présidentielle, est abordé. Les noms des candidats Chirac, Le Pen et Jospin sont particulièrement fréquents. Le thème de l'absence de choix figure dans ces énoncés, « *on n'avait plus le*

choix », « *pas assez de candidats au second tour* ». Ce discours est une représentation de l'événement du 21 avril 2002.

3.2.3. *Le ras le bol qui pousse vers une volonté de changement (classe 3)*

La volonté de changement est clairement affirmée dans les énoncés de cette classe. Les mots ou expressions « *changer* », « *changement* », « *alternance* », « *ras-le-bol de la gauche* », « *manque de confiance* » envers la gauche, « *manque d'information* » (délivrée par la gauche au pouvoir), « *promesses non tenues* » *(*du gouvernement de gauche) sont spécifiques de ces textes. On peut remarquer la fréquence élevée du mot Français : « *les Français aiment changer* », « *volonté de changement des Français* »,… Les extraits de discours de cette classe dénotent un point de vue très global.

3.2.4. *La politique de gauche a déçu (classe 4)*

Le thème principal de cette classe est la déception devant les actions du gouvernement de la gauche. Les problèmes relatifs aux « 35 heures », à l'immigration, aux inégalités sociales, à la sécurité, à la délinquance, à l'école, aux chômage, aux cotisations sociales, aux impôts, aux retraites… sont développés. Les jugements des enquêtés à l'égard de la gauche portent sur le manque d'écoute et de compréhension de la population, l'éloignement du terrain, l'incapacité à résoudre les problèmes, le fait qu'elle ne s'est pas montrée à la hauteur des attentes, ses erreurs, ses bêtises ...

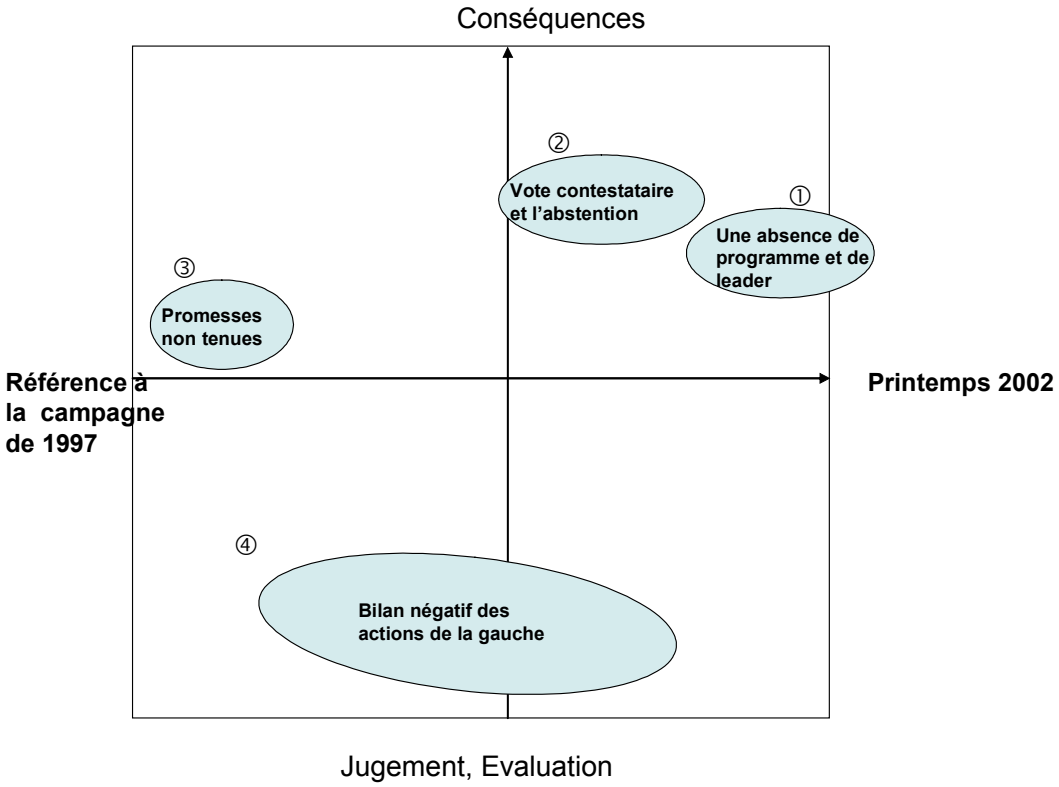
3.3. *Bilan de l'analyse lexicale*

L'analyse lexicale conduit à une typologie d'un certain nombre de discours tenus par les répondants sollicités à s'exprimer sur les résultats des élections de 2002. Certains types sont récurrents quelle que soit la formulation de la question. Il s'agit des énoncés concernant le bilan des actions de la gauche lors des cinq dernières années, la référence aux campagnes électorales jugées meilleures pour la droite que pour la gauche. En revanche, les résultats de l'analyse lexicométrique mettent en exergue des thèmes plus spécifiques de la victoire de la droite : le refus de la cohabitation, la peur de l'extrême droite et la volonté de changement, et d'autres spécifiques de la défaite de la gauche : des promesses non tenues, un vote contestataire et l'abstention.

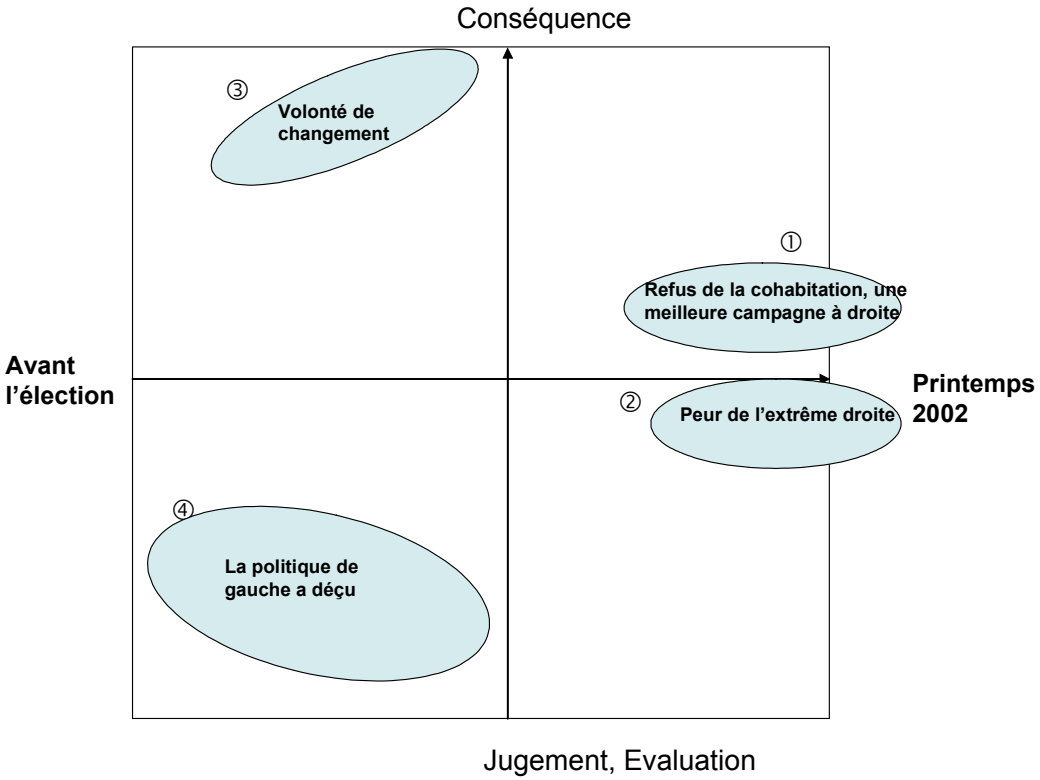
Les graphes présentés ci-dessus fournissent une image synthétique des premiers plans de l'analyse factorielle du tableau croisant le vocabulaire et les classes d'énoncés dégagées par le logiciel Alceste.

Le premier axe de chacune de ces deux analyses des correspondances peut être considéré comme un axe temporel. Dans le cas des réponses au questionnement sur les causes de la défaite de la gauche, les énoncés faisant référence à l'ancienne campagne s'opposent à ceux émettant un avis sur les faits du printemps 2002. Pour les réponses relatives à la victoire de la droite, des commentaires sur la situation politique à la veille des élections sont opposés à d'autres concernant les élections. L'intervalle de temps semble cependant plus étendu pour la défaite de la gauche que pour la victoire de la droite. Dans les deux cas, le deuxième axe de ces plans factoriels oppose des propos relatifs à une évaluation des actions de la gauche aux conséquences que ces actions ont eues sur les résultats du scrutin.

Figure 1. Représentation factorielle des classes Alceste
« Raisons de la défaite » (inertie : axe 1 : 37.8%, axe 2 : 34.7%)



« Raisons de la victoire » (inertie : axe 1 : 41.7%, axe 2 : 32%)



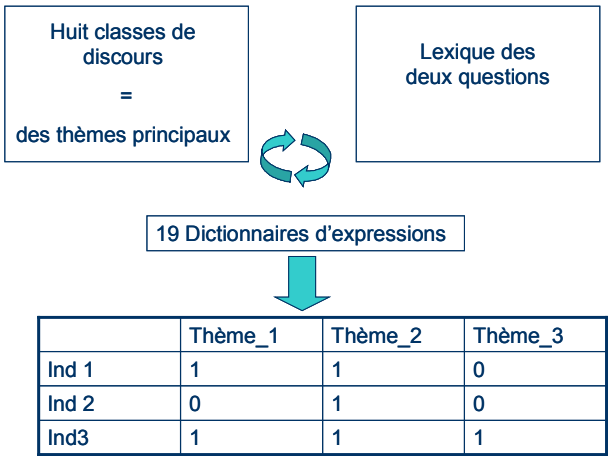
À partir de ces classes de discours, il convient maintenant, de mesurer l'influence de la question posée sur les thèmes abordés et de connaître le profil sociodémographique, les opinions et comportements politiques des individus tenant ces propos. Or, l'ensemble des verbatims n'est pas classé dans l'analyse précédente du fait de la grande variabilité de l'information recueillie et, dès lors, une étude statistique concernant l'ensemble des locuteurs s'avère délicate. De plus, les contextes dégagés précédemment sont parfois très hétérogènes (c'est le cas de la classe : « vote contestataire et l'abstention »). Ces deux raisons nous ont conduit à recourir à une analyse thématique pour affiner les classes de discours et revenir au plus près du répondant.

4. L'analyse thématique pour revenir aux répondants et à leur profil

4.1. *La démarche*

Chacune des classes de discours produite par Alceste est caractérisée par un ensemble de mots spécifiques. À partir de ces mots, des champs lexicaux sont isolés. Ces différents champs lexicaux constituent un premier niveau de thématique. À partir des champs lexicaux relatifs à un thème, les verbatims de réponses des individus sont explorés de façon à constituer un dictionnaire de mots et d'expressions associés à ce thème. La dernière étape consiste à créer, pour tout thème, une variable binaire prenant la valeur 1 lorsque l'enquêté évoque ce thème et 0 sinon. Cette phase de l'analyse a été facilitée par l'utilisation du logiciel Sphinx. Par cette démarche qui a nécessité un va et vient entre les classes d'énoncés, le lexique (ou les segments répétés) et leur contexte d'utilisation, nous avons affiné les thématiques. Par exemple, la classe de discours « bilan négatif des actions de la gauche » a été divisée en cinq sous-thèmes. Par ailleurs, cet approfondissement a permis d'observer l'importance de la thématique « je ne sais pas » non extraite lors de l'analyse lexicale précédente.

Figure 2. Organisation de la démarche



	Thème_1	Thème_2	Thème_3
Ind 1	1	1	0
Ind 2	0	1	0
Ind3	1	1	1

4.2. *Les thèmes et leur fréquence*

Dix-neuf thèmes ont été retenus. Certains, trop peu fréquents (expression du manque d'intérêt suscité par la question, « je ne m'intéresse pas à la politique ») n'ont pas été retenus. Dans 9% des cas, la réponse d'un individu n'est affectée à aucun des thèmes retenus, car aucun des mots ou expressions ne correspond aux dictionnaires thématiques créés. Ce sont pour la plupart des réponses peu claires ou trop ambiguës.

Les thèmes abordés dans les réponses ont été regroupés en six catégories :

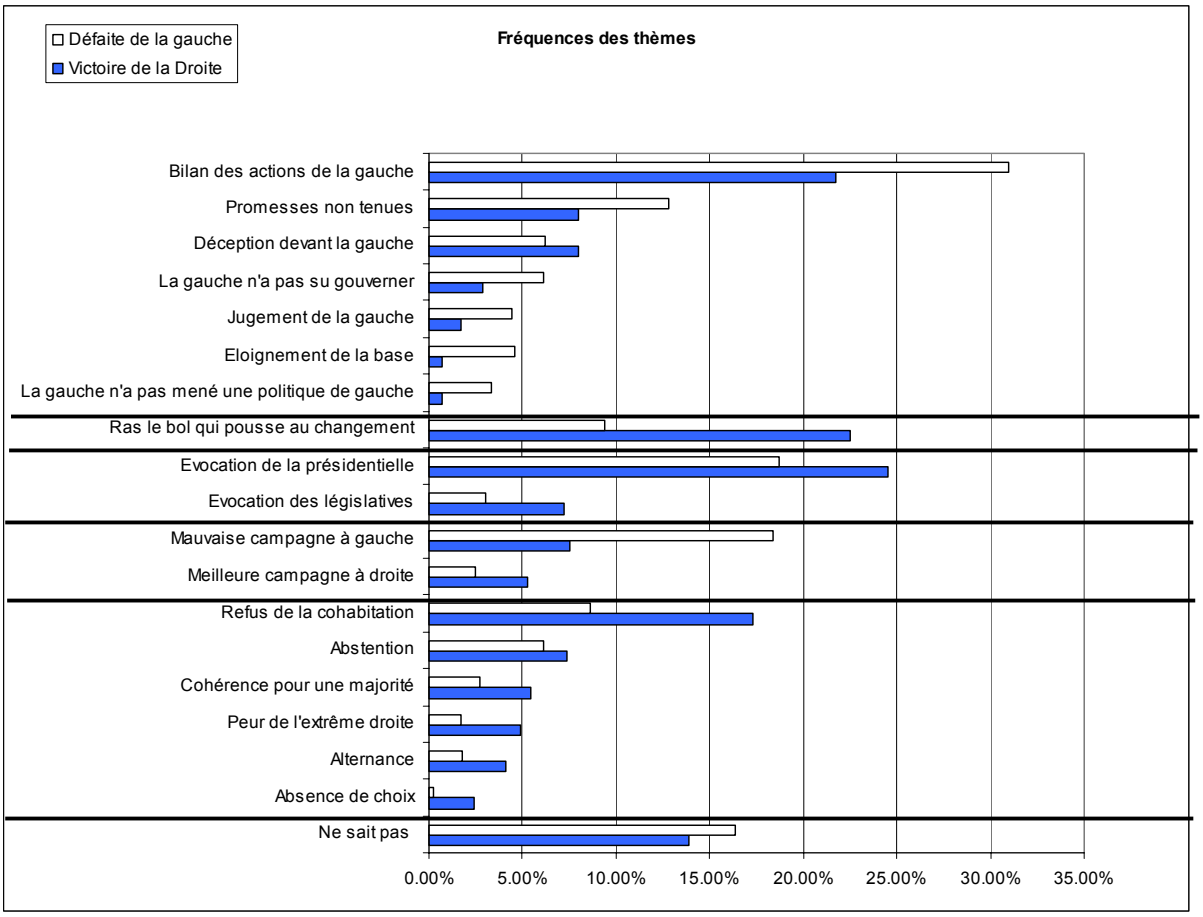
- *Le jugement sur la politique et les personnalités de gauche* constitue une première catégorie : le bilan du gouvernement de gauche, les promesses non tenues, le fait que la gauche n'a pas su gouverner, que la gauche n'a pas mené une politique de gauche, son attitude vis-à-vis de la base (le peuple, les électeurs, les Français). Le bilan négatif des actions de gauche, qui couvre entre autres les problèmes relatifs aux 35 heures, est un thème largement évoqué (26,2% des personnes interrogées). Il faut noter aussi l'importance de la thématique des promesses non tenues (9,2%).
- L'expression du *ras-le-bol de la gauche qui pousse au changement* est plus particulièrement rencontrée dans le corpus des raisons de la victoire de la droite (15,9% des personnes interrogées). Ce n'est pas la droite qui a gagné, c'est la gauche qui a perdu.
- La *référence à l'élection présidentielle* est très importante (21,6% des personnes interrogées). Ce thème est plus présent que celui des élections législatives (7,3%) Cette différence permet de penser que l'élection présidentielle est la référence lorsque l'on évoque cet ensemble d'élections. Les législatives ont été mises en sourdine.
- *L'évocation de la campagne de la gauche, jugée mauvaise*, est plus fréquente dans le texte des réponses associées aux raisons de la défaite de la gauche (12,8%). Le thème relatif à la *campagne de droite, jugée meilleure*, est plus fréquemment cité pour expliquer les raisons de la victoire de la droite. Il constitue cependant un thème très peu fréquent (3,8%).
- Les thèmes visant à expliquer les enchaînements logiques sous-jacents au déroulement des élections et faisant référence au contexte électoral d'après le 21 avril 2002 **,** *(l'abstention des électeurs - 7,5%, la peur de l'extrême droite - 5,2% , l'absence de choix au second tour de l'élection présidentielle - 1,3% , le refus de la cohabitation 12,9%, l'alternance - 3,3%, donner une majorité présidentielle - 3,4%)* sont plus largement évoqués dans le corpus des enquêtés interrogés sur les raisons de la victoire de la droite.
- La proportion des personnes exprimant leur embarras à expliquer les résultats des élections est importante (14,9%). Cela est vrai quelle que soit la question posée. Parfois, la réponse se limite à ces quatre mots « *je ne sais pas* » et dans d'autres cas, elle est ponctuée par cette expression.

Les relations entre les thèmes et la formulation de la question sont toutes significatives à l'exception des thèmes relatifs à « la déception devant la gauche », « l'abstention » et aux « je ne sait pas ».

Les critiques relatives à la politique menée par la gauche, au comportement des responsables politiques de gauche et à la campagne de la gauche sont plus particulièrement présentes dans le corpus des réponses analysant les raisons de la défaite de la gauche : ce résultat était attendu.

Pour ce qui concerne les causes de la victoire de la droite, le contenu du programme de la droite est peu commenté, les principaux arguments avancés par les locuteurs sont relatifs à l'articulation de l'ensemble des événements de la séquence électorale (le rôle de l'abstention, la peur de l'extrême droite, l'absence de choix). Ces éléments ont marqué les campagnes électorales et ont déclenché certains processus électoraux.

Figure 3. Les thèmes et leur fréquence en fonction de la formulation de la question



4.3. *Le retour aux répondants*

L'analyse a permis d'améliorer l'homogénéité des thèmes et d'associer une liste de thèmes à chaque répondant. La matrice individus-thèmes est l'outil avec lequel l'analyse sociodémographique et comportementale a été menée. Les thèmes les plus fréquents ont été conservés : le bilan des actions de la gauche, l'évocation de la présidentielle et les « ne sait pas ». Ces trois dimensions ont été croisées avec l'âge, le sexe, le diplôme, l'intérêt pour la politique, le positionnement politique et la reconstitution du vote du premier tour de la présidentielle. Ces liaisons ont été étudiées en contrôlant la formulation de la question.

Le premier constat que nous pouvons faire est que des liaisons existent mais sont dans la majorité des cas de faible intensité (V de Cramer< 0.2).

4.3.1. *Le profil sociodémographique des thématiques*

Le discours des femmes s'oppose à celui des hommes. L'absence d'opinion politique est plus fréquente chez les femmes. En revanche, les hommes sont plus nombreux à énumérer et à critiquer les actions de la gauche. Cette énumération est précise et témoigne d'un degré de connaissance supérieur de la vie politique. De plus, les hommes sont les plus nombreux à évoquer l'élection présidentielle. Cette situation confirme les résultats d'autres études selon lesquelles le niveau de politisation des hommes est supérieur à celui des femmes.

Dans le cas du niveau d'études, on a pu noter que le fait d'avoir un diplôme, le plus modeste soit-il, semble favoriser l'expression du discours. Les « sans diplômes » ont moins de certitudes pour expliquer l'issue des élections. Les plus diplômés sont plus nombreux à citer et commenter les actions de la gauche.

Les verbatims des moins de 25 ans révèlent que ce sont eux qui ont le plus de difficulté à expliquer ces événements, à l'instar des plus de 65 ans. À l'inverse ce manque d'argumentation se retrouve rarement chez les personnes âgées de 50 à 65 ans. Ce résultat confirme que la politisation et l'argumentation politique atteint un optimum dans cette classe d'âge.

L'ensemble des relations qui viennent d'être décrites sont les mêmes, que l'interviewé soit sollicité sur les motifs de la victoire de la droite ou sur ceux de la défaite de la gauche.

4.3.2. *Les opinions, attitudes et comportements politiques*

Le sens de la liaison entre les thèmes et l'intérêt pour la politique reste le même quelle que soit la formulation de la question. Les moins intéressés par la politique sont les plus nombreux à déclarer ne pas savoir pourquoi la droite a gagné ou la gauche a perdu. Ceux qui font le bilan des actions de la gauche lors des cinq dernières années ou ceux qui évoquent la présidentielle sont les interviewés qui se déclarent les plus intéressés par la politique.

En revanche, la liaison entre l'expression du vote au premier tour et la difficulté à trouver des raisons pour expliquer l'aboutissement de ces élections est plus forte lorsque l'on parle de la victoire de la droite. En effet, lorsque la question concerne la victoire de la droite, ce sont les extrémistes de droite ou de gauche qui déclarent ne pas savoir. Ce phénomène est beaucoup moins fréquent dans le discours des électeurs de droite. Ces électeurs sont moins hésitants, ils peuvent a minima exprimer les raisons qui ont été les leurs pour voter à droite.

Question posée	Présence du thème…	Extrême gauche	Gauche	Droite	FN	Total
Pourquoi la gauche a-t-elle perdu ?	**« Ne sait pas »**	20,4%	12,0%	13,1%	14,9%	13,8%
	Effectif Total	103	325	290	74	792
Pourquoi la droite a-t-elle gagné ?	**« Ne sait pas »**	17,3%	12,7%	8,1%	20,0%	12,4%
	Effectif Total	110	315	310	90	825

Fréquence du thème « Ne sait pas » selon le vote au premier tour de la présidentielle

5. Conclusion

Si l'analyse du discours a conduit à l'élaboration de huit types de discours, l'analyse thématique a permis d'affiner cette typologie en dégageant dix-neuf thèmes.

Cet enchaînement méthodologique offre la possibilité de mesurer l'importance des thèmes présents dans l'opinion publique pour expliquer l'issue de la séquence électorale de 2002. Ainsi l'évocation de l'élection présidentielle est prépondérante sur l'évocation des élections législatives. Le bilan des élections de 2002 semble majoritairement associé aux résultats de l'élection présidentielle. Les élections de référence ne sont pas les mêmes pour expliquer la victoire ou la défaite, tout découle du premier tour de la présidentielle et tout est en place au soir du 21 avril 2002.

Le deuxième point fort de cette démarche est le retour aux locuteurs caractérisés par l'ensemble des thèmes qu'ils ont énoncés. Ainsi nous pouvons éclairer les motifs présents dans les réponses par le profil sociodémographique, l'opinion et le comportement politique des répondants.

Après une première lecture des textes recueillis, les énoncés semblaient lapidaires, peu argumentés et comportant des informations très diffuses. Cet état des lieux laissait présager des difficultés à faire émerger une structure dans ce corpus. En fait, ce matériau s'est avéré très riche et structuré. En effet, il a fait apparaître que les raisons de la défaite font référence au programme que la gauche a proposé aux législatives de 1997, aux cinq années de politique de gauche, à la campagne du premier tour de la présidentielle et aux résultats du 21 avril 2002. Par ailleurs, les raisons de la victoire de la droite renvoient aux actions de la gauche jugées de manière négative, à la situation politique issue de l'état de choc qui a suivi le 21 avril 2002. Cette structuration du corpus permet de confirmer l'expression d'un « vote de tous les refus » (Perrineau et Ysmal, 2002) : un bilan négatif de la gauche qui conduit à son désaveu, un rejet de la vie politique par une abstention en forte hausse, un refus de l'extrême droite à l'issue du premier tour de la présidentielle et un rejet de la cohabitation, expérience jugée non convaincante. Ne s'agirait-il même pas d'un vote de tous les rejets ? .

Références

Le Panel Electoral Français 2002. site web : http ://cidsp.upmf-grenoble.fr/index_postelec.htm

Benzecri J.-P. *et al.* (1981). Pratique de l'analyse des données, vol. (3). Dunod.

Blot I., Hammer B. et Le roux D. (1994). Traitement des questions d'opinion « ouvertes » : utilisation d'Alceste, outil d'assistance à l'analyse. *Revue ICO Québec*, vol. (6/1-2).

Brugidou M. (1998). Épitaphes, l'image de François Mitterrand à travers l'analyse d'une question ouverte posée à sa mort. *Revue Française de Science Politique*, vol. (48/1) : 97-120.

Brugidou M. (2001). La combinaison des inférences statistiques, linguistiques et sociologiques dans l'analyse d'une question ouverte. *Journal de la Société Française de Statistique*, vol (142/4).

Lebart L. et Salem A. (1994). Statistique Textuelle. Dunod.

Muller Ch. (1993). Principes et méthodes de statistique lexicale. Slatkine-Champion.

Perrineau P. et Ysmal C. (2002). Le vote de tous les refus. Presses de Sciences Po.

Reinert M. (1987). Classification descendante hiérarchique et analyse lexicale par contexte : application au corpus des poésies d'Arthur Rimbaud. *Bulletin de Méthodologie Sociologique*, vol. (13).

Le rapport à l'autre dans la psychose bipolaire

Sylviane Burner

UFR Lettres et Langues – Université – 57045 Metz Cédex – France
sylviane.burner@wanadoo.fr

Abstract

Manic-depressive psychosis is characterised by two alternative periods of depression and mania with remissions in-between. Each state has specific features. As in all psychoses, the relation between the patient and the outside world is altered. It seemed interesting to study the way these persons use the 1st and 2nd person forms, to see how the "I" relates to the "you". The corpus is in English, made of 75 709 words obtained through non directive interviews with the patients, recorded then transcribed. The results show that the 1st person system is used in opposite ways in depressive and manic phases, the pronouns being mainly used as grammatical subject in the depressive period, while they are complement in the manic phase. This finding comes as contradictory to the superficial visible characteristics of the illness. The analysis of the texts produced during a remission shows that the self is still fragile and unable to assert itself. The use of the 2nd person seems more or less similar in the three states. Even if the interviewer is perceived and addressed to as such, most of the forms are included in phatic or generic forms. This study shows that the remission phase is closer to a period in which symptoms abate rather than disappear, and points out the fact that although the patient seems on top of the world in the manic phase, an underlying depression might still be very present in the manic state.

Résumé

La psychose bipolaire fait alterner des phases de dépression et des phases maniaques, entrecoupées de périodes de rémission pendant lesquelles le sujet présente des particularités spécifiques. Comme dans toute psychose, c'est le rapport à soi et au monde qui est perturbé. Il nous a semblé intéressant d'étudier le système des formes de 1re et de 2e personnes au cours de ces trois phases, autrement dit le rapport du « je » au « tu/vous ». Le corpus analysé est en langue anglaise et rassemble 75 709 mots obtenus grâce à des entrevues libres enregistrées puis transcrites. Il apparaît que les formes de 1re personne sont utilisées de façon presque opposées en phase dépressive et en phase maniaque, l'une privilégiant la position « sujet », l'autre « complément », allant ainsi à l'encontre de toutes les idées reçues sur cette pathologie. Qui plus est, la phase de rémission semble traduire un rapport à l'autre encore très fragile. Dans les trois sous-corpus, l'utilisation de la 2e personne présente des similitudes quant à la pauvreté de l'adresse réelle à l'interlocuteur et quant à l'importance des formes phatiques et des formes génériques. Cette étude laisse à penser que la phase de rémission est plus proche d'une période de contrôle des symptômes que d'une guérison, et que l'état maniaque, pendant lequel le sujet exprime son bien être et son optimisme, garde une forte composante dépressive sous-jacente, masquée certes, mais bien réelle.

Mots-clés : Psychose bipolaire, dépression, manie, formes 1re et 2e personnes, rapport à l'autre.

1. Introduction

La linguistique appliquée à la psychiatrie est une discipline récente. Les premiers écrits datent des années 70, et il faut avouer que l'essor de cette science reste lent. Une grande partie des travaux porte sur la thématique des discours, qu'ils analysent souvent dans une perspective freudienne ou lacanienne. À l'inverse du psychanalyste, le linguiste s'intéresse plus particulièrement à l'agencement des représentations et moins à leur contenu. Il s'attache alors aux marqueurs linguistiques susceptibles de varier en fonction de l'état psychique de la personne. Les travaux de Tonus (1993) montrent que la négation et les verbes statifs semblent caractériser les états dépressifs. La diversité lexicale et la modification de la relation pauses-

parole seraient, par contre, le fait des états maniaques (Burner, 1980a). Mais quelle que soit la pathologie, c'est prioritairement le rapport au monde et aux autres qui est affecté, comme en témoigne l'emploi des pronoms personnels (Burner, 1985 et 1993 ; Reb et Trognon, 1986). Il est à remarquer que la plupart des analyses se rapprochent plus de l'étude de cas que de l'analyse d'un corpus de taille respectable, ce qui limite considérablement les possibilités de généralisation.

Le travail proposé ici se concentre sur la psychose bipolaire et vise à établir les variations du rapport au monde et à l'autre au cours des trois phases de la maladie : dépression, manie et stabilisation. Pour ce faire, c'est l'analyse des formes grammaticales de première et deuxième personnes qui a été retenue et effectuée sur un corpus de grande taille.

2. La psychose bipolaire

La psychose bipolaire est une forme de psychose maniaco-dépressive, pathologie évoquée pour la première fois par Falret dans *De la folie circulaire ou forme de maladie mentale caractérisée par alternative régulière de la manie et de la mélancolie* (Falret, 1851). En 1854, Jules Baillarger, la décrit une « une folie à double forme » (Baillarger, 1854). À leur suite, de nombreux médecins et chercheurs ont affiné ces définitions et montré l'extrême complexité et la grande diversité des formes de cette pathologie.

Le système de classification des pathologies mentales qui a été retenu pour cette étude est le DSM IV[1], reconnu par l'OMS[2], qui est régulièrement mis à jour et qui sert de référence en clinique, en épidémiologie et en pharmacologie.

La psychose bipolaire se caractérise par la survenue, chez un même sujet, d'épisodes de dépression majeure et d'épisodes d'excitation maniaque. Ces différents accès sont séparés par des phases de rémission de plus ou moins bonne qualité. La réapparition de l'un ou de l'autre état pathologique peut être progressif ou brutal.

La phase dépressive présente les caractéristiques connues de la dépression (repli sur soi, tristesse, insomnie, lenteur, fatigue, etc.). L'accès maniaque présente une logorrhée, une exaltation de l'humeur, une accélération dans tous les domaines d'activité : l'appétit et la soif augmente, le désir sexuel aussi, et l'insomnie est presque totale.

Si le sujet dépressif est triste, taciturne et peu communicatif, le sujet maniaque est omniprésent, débordant d'activité, noyant ses proches sous un flot de paroles, se sentant heureux et maître du monde.

Ce bonheur affiché, difficilement compatible avec la maladie, m'a amenée à m'interroger sur sa véracité. Peut-on réellement passer, parfois en très peu de temps, d'un état suicidaire à un état de jubilation extrême — et inversement —, en gommant toutes les caractéristiques de la pathologie précédente ?

J'ai donc entrepris une analyse des discours de patients présentant une psychose bipolaire avec phases de rémission.

[1] **D**iagnosis and **S**tatistical manual of **M**ental **D**isorders.
[2] **O**rganisation **M**ondiale de la **S**anté.

3. Le corpus

Le corpus a été réuni par mes soins en milieu hospitalier, en France et en Angleterre. Il se compose d'entretiens réalisés selon un protocole bien précis dans lequel les patients sont amenés à parler de façon libre, avec une intervention minimum de la part de l'interlocuteur (Burner, 1980b). Les entretiens enregistrés sont ensuite transcrits. Si le recueil des données est relativement facile en phase maniaque, il est souvent plus laborieux en phase dépressive. J'ai néanmoins réussi à réunir un nombre suffisant de textes pour une analyse à grande échelle. L'ensemble du corpus concernant la psychose bipolaire se compose de **75 509 mots**, se décomposant en **25 452 mots** pour la phase maniaque, **25 048 mots** pour la phase dépressive et **25 009 mots** pour la phase de rémission.

L'étude menée ici concerne exclusivement les textes oraux en langue anglaise et a été traité grâce aux logiciels *Le Concordeur 2.0* (Rand, 1991) et *Hyperbase* 5.4 (Brunet, 2001).

4. Les formes 1re personne

Outre le refus de se considérer comme malade, une des caractéristiques de la psychose est une altération du rapport à l'autre et au monde, une indistinction entre le « moi » et le « non-moi ».

Dans cette optique, l'étude des formes grammaticales de 1re personne (*I, me, my, myself, we, us, our, ourselves*), et de 2e personne (*you, your yourself),* les seules véritablement concernées par l'acte d'énonciation (Benveniste, 1966), m'a semblée intéressante à plusieurs titres. Je ne pose ici pas l'hypothèse que ces marques linguistiques sont les seules pertinentes pour évaluer les rapports de la personne, mais il est indéniable qu'elles en constituent des indices directs.

Avant d'entamer une étude linguistique du contexte d'apparition des formes de 1re et 2e personnes, il m'a paru utile de procéder à un recensement brut de l'apparition de ces formes dans les discours étudiés. Le tableau 1 qui les regroupe permet déjà de percevoir des modifications en fonction des trois phases de la psychose bipolaire.

Phases	**1re pers. sg (%)**	**1re pers. pl. (%)**	**1repers. pl. (%)**	**2e pers. (%)**
Manie	8.88	0.90	9.78	7.46
Dépression	9.14	1.29	10.43	8.08
Rémission	1.33	0.89	2.22	13.45

Pourcentages des formes 1re et 2e personnes selon les 3 phases étudiées

Le chi-2 effectué sur les effectifs bruts est de 31.53 pour un degré de liberté de 6, ce qui montre qu'il n'y a qu'une chance sur 1000 pour que la répartition de ces formes dans les textes soit due au hasard (Siegel, 1956).

Les chiffres montrent d'emblée une relative similitude d'emploi de ces formes en phase maniaque et en phase dépressive, et une modification importante en phase de rémission. Il semblerait donc, au premier abord, que la relation à l'autre ne soit guère différente dans les deux phases pathologiques. Il est également intéressant de remarquer que les formes de 1re personne plurielle sont peu représentées, suggérant ainsi une difficulté du sujet à s'inclure dans la relation même avec l'autre.

4.1. *Textes produits en phase de rémission*

Si le premier constat qui s'impose chez les sujets stabilisés est le recul des formes de 1re personne par rapport aux phases pathologiques, l'analyse en contexte permet de voir que la forme singulier (***I, my***) est souvent suivie d'une forme plurielle (***we, our***) :

(a)…the most satisfying period of my career came during the war many people had no houses we won the battle of Britain…

(b)…I went abroad and I know that in our present day polygamy has been practiced in certain sects…

Le "I" qui raconte semble garder ses distances avec les faits racontés en utilisant, à la suite de ce pronom, un pronom 1ère personne pluriel générique, éprouvant le besoin de

«donner à « nous » une compréhension indéfinie et l'affirmation volontairement vague d'un « je » prudemment généralisé » (Benveniste,1966).

Le même phénomène est décrit par Vion (1992) sous le terme « d'effacement énonciatif ».

On peut d'ailleurs remarquer, dans l'environnement proche de ces formes « dilatées » l'abondance de tournures impersonnelles et passives, de formes indéfinies :

« they say in some cases, this happen » ;

« one says that one can tell you in the what they call the Council House… »,

comme si le malade n'avait pas encore suffisamment repris confiance en lui pour affirmer sa parole.

La thématique, par ailleurs, privilégie les thèmes généraux et évoquent fréquemment les collectivités (the government, the Council House, mankind), permettant ainsi au locuteur de se fondre dans l'anonymat du groupe. Ainsi,

« le je s'amplifie par nous en une personne plus massive, plus solennelle et moins définie » (Benveniste,1966).

Le « je » retrouvé après les crises de la psychose est donc un « je » fragile, hésitant, qui n'ose pas encore s'affirmer dans le « nous ».

Comme le montre le tableau ci-dessus, le « je » est par contre étonnamment présent, voire omniprésent, dans les discours recueillis en phases dépressive et maniaque.

L'étude détaillée des formes de 1re personne utilisées au cours de ces deux phases montre des constantes et des différences intéressantes.

4.2. *Textes produits en phase dépressive*

En phase dépressive, le « je » énonciateur envahit le discours. L'analyse des concordances montrent qu'il apparaît essentiellement en position sujet, donc sous la forme « I », très souvent renforcée par la présence de « my », forme d'appropriation souvent associée au thème de la parenté :

*« **my husband** is a teacher and **I** cannot manage on his wages **I** was well back at work with **my friends** and then looking after **my mother** and **father** on the days that **I** could if I have to give **my** job up because of **my parents**…".*

Le « I » est par ailleurs très souvent redoublé :

« ***I I my husband*** *is older than* ***me*** *»,*

*« well **I I** am **I** am about and **I** believe **I** wish **I** wish this government… »"*

Par cette abondance de "I", proche du tic de langage, le locuteur dépressif arrive ainsi à marteler sa présence, à affirmer son maintien d'une phrase à l'autre, à en faire un élément déterminant de la conduite de son discours.

Quant à l'environnement verbal de ce pronom, il est essentiellement constitué de verbes renvoyant à un état ou une caractéristique « *I am /I was*», à des processus mentaux « *I think », « I believe »,* ou à des sentiments *« I felt, I enjoyed, I like »*, très souvent à la forme négative (Burner, 2004).

Tout se passe en fait comme si le locuteur, par la surabondance de ces formes verbales, voulait affirmer la réalité et la pertinence de sa trame narrative. Le discours joue sur l'excès en émaillant le récit d'une profusion d'avis personnels factices, car ne se rapportant qu'à des formes vides de sens. De plus, les évènements évoqués ne sont pas racontés, mais simplement énumérés, juxtaposés. Les principes qui les rassemblent ne sont pas des liens de causalité, ni des liens de succession logique mais des lois régies par l'environnement et la mémoire, comme l'a par ailleurs observé Ghiglione qui note que chez les dépressifs,

« Les univers référentiels observés semblent concerner le rôle accordé à la famille et à la mémoire » (Ghiglione R *et al.*, 1995).

Si les chiffres montrent que les formes de 1re personne sont également fortement représentées chez les malades en phase maniaque, l'analyse de détail laisse apparaître un fonctionnement autre.

4.3. *Textes produits en phase maniaque*

Contrairement à leur utilisation en phase dépressive, les formes de 1re personne, au cours de la phase maniaque, apparaissent prioritairement en position complément :

« it bores **me** to tears the sort of physics did **my my** research the things that interest **me** now seem more basic » ;

« there is nothing for **me** to go as a day patient nothing for **me** there for **me** to do".

Les pronoms personnels sujet sont, quant à eux associés à des verbes de volition (to want) essentiellement ou des auxiliaires de modalité (can, could) comme on peut le voir dans les exemples suivants :

« all **I want** to do is **I want** to go out go dancing see people but they could not » ;

«**I can** touch people but **I can't** say now that **I**'ve got is something got to do with God **I can't** have anything wrong mentally » ;

« **I can** sing wonderful but **I won't be able** to sing any more not if they don't hurry up".

Si l'association du **I** et de l'évocation du pouvoir et du vouloir rentre parfaitement dans le tableau que l'on brosse généralement des patients maniaques, il est paradoxal de voir que l'utilisation principale des formes premières personnes est la forme complément. Il semble que tout en affirmant sémantiquement sa toute-puissance, la personne avoue son assujettissement à la maladie, trahisse le fait qu'elle subit sa vie sans en être l'acteur qu'elle voudrait être. Les derniers exemples sont intéressants à ce point de vue puisqu'on peut remarquer que l'affirmation du vouloir et du pouvoir (**I want ; I can**) est immédiatement suivie par un démenti (**but they could not**), démenti ici d'autant plus fort qu'il est imposé par « **they** » ou

par la prise de conscience que ce pouvoir ou de vouloir n'est que virtuel (**I can't ; I won't be able to**).

L'affirmation sémantique du bonheur ("**I** never sit still because **I**'m happy and **I** like to make babies") ne serait-elle qu'un masque destiné à adoucir le total désarroi de ces patients? Le maniaque serait-il, à l'instar du dépressif, profondément triste et malheureux, en dépit des apparences ?

5. Les formes 2e personne

5.1. *Un interlocuteur réel*

L'utilisation des formes de 2e personne donnera peut-être des indications sur le rapport que le maniaco-dépressif entretient non seulement avec le monde, mais avec ses semblables, en situation de communication.

Le tableau ci-dessus montre que les formes 2e personne sont présentes dans nos trois sous-corpus, et augmentent de façon relativement importante en phase de rémission, suggérant peut-être alors le rétablissement du dialogue avec l'autre.

Il est intéressant de noter que les particularités de l'emploi des formes de 2e personne comportent des constantes que l'on retrouve dans les trois phases.

En effet, même si la présence de l'interlocuteur semble toujours perçue comme réelle, la valeur dialogique est faible. Les formes d'adresse directe à l'autre restent ainsi organisées autour de la fonction sociale de la personne :

« ***you*** *are now in* ***recording this tape you*** *got to justify your wages » ;*

*« **I** can talk to **you** because **you** are a **doctor » ; « you** are **treating me**".*

Cette particularité est encore plus marquée dans le cadre du discours maniaque.

Quant à la phase dépressive, elle présente une particularité qui montre aussi l'impossibilité pour le locuteur d'assumer sa place. En effet, on trouve des énoncés qui montrent que le locuteur refuse de figurer comme tel dans son discours :

« ***we*** *got married and it did take* ***you*** *some time to adjust to one another » ;*

*« **in this case I** do think **you** enjoy it because **you** are not used to* ».

Dans ces cas, le "*I*" se découvre, puis se réfugie aussitôt derrière un « *you* », refusant de figurer comme actant dans ses propres énoncés.

5.2. *Les formes phatiques*

D'autre part, la fonction phatique du langage (Jakobson, 1973) est extrêmement présente dans ces formes de 2e personne :

« *I'm a hypomaniac* ***you know you see*** *» ;*

« I got on the bus and the people all seemed jazzy clothes ***you know*** *they've got* ***you know*** *and every time I see this you see it makes me wonder…".*

L'adresse à l'autre est ici une façon de préserver le contact, en aucun cas d'inclure l'interlocuteur dans l'échange verbal. Ces formes vides de sens, caractéristiques des discours psychotiques (Burner, 1987) se retrouvent dans tous les textes étudiés, à des degrés divers et quelle que soit la phase considérée. Le désir de contact semble toujours réel puisque le locuteur fait l'effort d'interpeller le co-locuteur, mais le dialogue reste fictif.

5.3. *Les formes génériques*

Cette particularité est commune à nos trois sous-corpus, tout comme l'est celle liée à l'utilisation du « you » générique :

« *when **you** make love, **you** should think of… » ;*

*« when **you** live **you** carry a certain amount of electricity » ;*

*"**you** never know what they can do to **you**".*

Cette fonction de "*you*" est extrêmement importante dans nos trois corpus, prioritaire dans les textes en phase de rémission, et met le doigt sur la stratégie de flou référentiel à laquelle les psychotiques ont recours pour communiquer tout en évitant de trop s'impliquer personnellement dans leur discours.

6. Conclusion

Au terme de cette étude, il apparaît que le rapport au monde et à l'autre est problématique dans toutes les phases de la psychose bipolaire. Le « je » sujet « vide », envahissant, de la phase dépressive cède la place au « moi » passif, subissant, dans la phase maniaque, alors même que le discours thématique est dominateur et optimiste, tandis que le « je » et le « moi » de la phase stabilisée se diluent dans le générique et le flou référentiel. La phase qui pouvait être vue comme « rémission » n'est guère plus qu'un gommage des excès des deux phases extrêmes de la maladie, un espace où le patient a encore bien du mal à trouver des repères, suggérant ainsi que le facteur dépressif est masqué plutôt qu'absent. Le « je » se cache derrière un « vous » non identifié qui se dilue dans le flou référentiel et par ce biais peut-être, se préserve. On peut émettre l'hypothèse que le facteur dépressif, nié par la thématique du discours, reste néanmoins très présent dans la phase maniaque.

Ces diverses stratégies discursives peuvent donc apparaître comme un recours habile au langage pour conserver un semblant de communication tout en tenant l'autre à distance, pour préserver le minimum vital de contact sans se sentir en danger. Il reste que cet « autre » que nous sommes aussi a encore beaucoup de chemin à faire pour comprendre et déchiffrer, comme le dit Thellier (1980),

« les destins rompus de ceux qui ont emprunté des routes qu'ils ne se sont pas choisies »

Références

Baillarger J. (1854). Folie à double forme. *Annales médico-psychologiques du système nerveux*. Masson.

Benveniste É. (1966). *Problèmes de linguistique générale*, t. (I). Gallimard.

Brunet Ét. (2001). Logiciel *Hyperbase* 5.4.

Burner S. (1980a). Influence du thème sur la production du discours oral. *Verbum*, vol. (3) : 37-54.

Burner S. (1980b). Répartition et fonction des temps de pause et de parole dans un cas de manie-dépressive. In *Psychologie médicale*, vol. (12/9) : 1867-1875.

Burner S. (1985) « L'autre » et le « moi » dans les délires psychotique. *Psychologie médicale*, vol. (13/3) : 57-63.

Burner S. (1987). *Étude du processus de rupture de communication dans les délires psychotiques*. Thèse de Doctorat d'Etat, Université de Paris VII.

Burner S. (1993). Émergence de facteurs dépressifs dans les délires psychotiques. In *Mélanges offerts à Véronique Huyn-Armanet : le texte : un objet d'étude interdisciplinaire*. Publications du LARIT : 2.

Burner S. (2004) Modification de l'affect dans les psychoses bipolaires. In *Mélanges offerts à Michel Morel*. Publications de l'université de Nancy. (A paraître).

Falret J.P. (1851). *Maladies mentales et asiles d'aliénés. Leçons cliniques et considérations générales avec un plan de l'asile d'Allenau*. Baillère.

Jakobson R. (1973). *Questions de poétique*. Seuil.

Ghiglione R *et al.* (1995). *L'analyse cognitivo-discursive*. PUF.

Rand D. (1991). Logiciel *Le concordeur* 2.0. Les publications CRM, Université de Montréal, Canada.

Reb V. et Trognon A. (1986). L'adhérence au discours de l'autre - (analyse pragmatique d'une conversation avec un psychotique). *Perspectives psychiatriques*, vol. (1).

Schultz V. (1999). *Étude du comportement discursif de sujets souffrant de psychose maniaco-dépressive : la triade énonciative*. Mémoire de maîtrise, Université de Metz.

Siegel S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. Mc Graw-Hill.

Thuillier J. (1980). *Les dix ans qui ont changé la folie*. R. Laffont.

Tonus F. (1993). *La dépression, la négation et le non-dit*. Mémoire de maîtrise, Université de Metz.

Vion R. (1992). *La communication verbale*. Hachette.

La percezione della sinonimia : un'analisi statistica mediante modelli per ranghi

Carmela Cappelli, Angela D'Elia

Dipartimento di Scienze Statistiche – Università di Napoli Federico II
Via L. Rodinò, 22 – 80138 Napoli – Italia
{carmela.cappelli ; angela.delia}@unina.it

Abstract

In this paper we deal with the role of synonyms in the Italian language, focussing on the way they are perceived by the people. In particular, we propose to exploit statistical models for ranks data in order to analyse the rankings expressed by different raters towards the set of synonyms of some different words. Indeed, by means of these models, we can highlight the level of perceived synonymy with respect to a "target" word and the presence of uncertainty in the ranking process itself ; moreover, we can study the existence of a link between the raters' covariates (e.g. sex, age, education level, etc.) and the ranks they give among the synonyms of a given list.

Riassunto

L'articolo si colloca nell'ambito degli studi sull'uso dei sinonimi nella lingua italiana, focalizzando l'attenzione sulla percezione – da parte dei parlanti – della sinonimia tra parole. In particolare, si propone il ricorso a modelli statistici per variabili rango al fine di analizzare le graduatorie che diversi soggetti formulano con riguardo alle liste di sinonimi di alcune parole. In effetti, mediante tali modelli per ranghi è possibile quantificare sia il livello di sinonimia percepita rispetto ad una parola "obiettivo", sia il grado di incertezza presente durante l'elaborazione della graduatoria stessa. Inoltre, è possibile analizzare il legame esistente tra le principali caratteristiche dei soggetti (come sesso, età, livello di istruzione, ecc.) e le graduatorie che essi esprimono.[1]

Keywords : synonyms, rankings, word senses' identification, MUB model.

1. Introduzione

Nel corso degli ultimi decenni si è manifestato un crescente interesse verso l'uso di metodi e modelli statistici per lo studio di problemi di natura linguistica, come testimoniato dai numerosi testi che trattano dell'impiego della statistica per l'analisi di dati testuali (si vedano tra gli altri : Woods *et al.*, 1986 ; Lebart *et al.*, 1998 ; Bolasco, 1999).

La linguistica, tradizionalmente, opera una distinzione tra la morfo-sintassi, che attiene alle regole che presiedono alla formazione delle frasi o delle parole, e la semantica, che studia invece il significato delle parole o delle frasi e, dunque, attiene al contenuto di un testo. In quest'ultimo ambito, si è sviluppata in tempi recenti una notevole attenzione per le problematiche relative alla similarità semantica : cioè, la sinonimia (Lin *et al.*, 2003 ; Ploux e Ji, 2003).

[1] Il presente lavoro è frutto di una comune ricerca degli Autori. C. Cappelli ha scritto i paragrafi 1, 3, 4.1 ; A. D'Elia ha scritto i paragrafi 2, 4.2, 5.

In tale contesto, il presente lavoro affronta –mediante un approccio modellistico– un particolare aspetto del tema della similarità semantica : *il grado di sinonimia tra parole, così come risulta essere percepito da parte degli utilizzatori di una lingua*.

In effetti, la conoscenza e l'uso corretto dei sinonimi rappresenta un indicatore importante della padronanza di una lingua. Tuttavia tale uso non è univoco, legandosi al problema, noto in semantica, della polisemia. Il termine polisemia, introdotto dal linguista francese M. Bréal nel 1897, sta ad indicare la complessità semantica di una parola, ovvero la coesistenza di più significati in una stessa parola : si pensi, ad esempio, al sostantivo *albero*, cui corrispondono i diversi significati di pianta (botanica), grafico genealogico (araldica), organo di acciaio per reggere le vele (nautica) o per trasmettere movimento alle ruote (meccanica).

Al fine di meglio inquadrare il concetto di polisemia, è opportuno riflettere sulle distinte nozioni di *significato* e *senso*. Come discusso in De Mauro (2002), il significato rappresenta l'insieme di tutti i valori ed usi che una parola può assumere nella lingua ; il senso, invece, è il modo in cui la parola è sentita (percepita) dalla persona che la utilizza e/o la deve interpretare : esso, quindi, identifica valori ed usi determinati e particolari. Mentre il significato appartiene alla lingua ed alla comunità dei suoi utilizzatori ed ha quindi natura comune, il senso riguarda l'esprimersi individuale : è il modo in cui il significato si estrinseca da parte di chi parla o scrive. Appare, dunque, evidente che i dizionari – ed in particolare i dizionari dei sinonimi – diano conto del significato delle parole, individuando delle accezioni, ossia sensi consolidati in quanto ripetuti e ripresi dagli utilizzatori della lingua. Per contro, per i sensi, in quanto aventi natura occasionale e soggettiva, è lecito ipotizzare che siano strettamente legati alle caratteristiche personali e siano anche frutto di un meccanismo di percezione da parte dell'individuo (in base alla sua cultura, alle sue esperienze, ecc.).

Oggetto del presente lavoro è lo studio dei sinonimi come possibili sensi di una parola "obiettivo" : tale studio, alla luce della problematica esposta, può essere condotto mediante l'analisi delle scelte individuali espresse in termini di graduatorie di similarità percepita. In particolare, se si assume che ciascuna parola sia caratterizzata da uno spazio semantico definibile attraverso l'insieme dei suoi sinonimi[2] (Cappelli, 2003 ; Cappelli e Corduas, 2003), l'analisi delle graduatorie elaborate dagli utenti sugli elementi facenti parte di tale spazio (ordinandoli in base al grado di sinonimia percepita) può essere efficacemente condotta mediante l'impiego di modelli statistici per ranghi. La proposta di utilizzo di tali modelli, in effetti, deriva dalla considerazione che essi consentono di quantificare il livello di sinonimia percepita e di individuare l'esistenza di sensi in base al legame tra le principali caratteristiche dei soggetti e le graduatorie da essi stessi espresse.

Il presente lavoro è così organizzato : nel paragrafo 2 vengono descritte, nel dettaglio, le motivazioni dell'approccio proposto e la metodologia sviluppata per l'analisi della percezione della sinonimia, mediante l'impiego di un modello mistura per variabili rango ; nel paragrafo 3 si illustrano le caratteristiche della indagine condotta per la valutazione della sinonimia percepita ; la presentazione ed il commento dei risultati della indagine costituiranno oggetto dei paragrafi 4.1 e 4.2. Alcune considerazioni finali concludono il lavoro.

[2] E' evidente che si tratta di una ipotesi di lavoro ; nulla vieta che se ne adottino altre. Ad esempio, che si definisca lo spazio semantico di una parola a partire dall'insieme dei suoi antonimi, oppure dei sinonimi e degli antonimi.

2. La metodologia

Si consideri una parola obiettivo w e sia $S_w = [s_1, \ldots, s_j, \ldots, s_m]$ l'insieme di tutti i suoi possibili sinonimi, individuati sia mediante ricerca manuale che elettronica (come esplicitato in Cappelli, 2003 ; Cappelli e Corduas, 2003).

Si assuma, inoltre, che n soggetti elaborino una graduatoria degli m elementi di S_w secondo un criterio di sinonimia rispetto a w : ogni "giudice", cioè, assegna rango R=1 al sinonimo (tra gli m a disposizione) che percepisce come più *simile* alla parola w, rango R=2 a quello che percepisce come successivo, e così via, fino ad arrivare al vocabolo che viene percepito come il più lontano in termini di sinonimia e che, quindi, riceve rango R=m.

In tal modo, ad ogni prefissato elemento s_j (j = 1, 2, …, m) di S_w è associato un vettore di ranghi osservati $\mathbf{r} = (r_1, r_2, \ldots, r_n)'$ che rappresentano una misura del grado di sinonimia tra s_j e la parola obiettivo w così come viene percepito dagli n giudici. Tale tipologia di dati può essere efficacemente analizzata mediante modelli per variabili rango (per una rassegna : Marden, 1995 ; D'Elia e Piccolo, 2002).

Al fine di proporre un modello statistico adeguato a rappresentare il meccanismo generatore dei ranghi osservati $\mathbf{r}$ per ogni s_j, è utile riflettere sulla procedura psico-linguistica che, presumibilmente, presiede all'elaborazione di una graduatoria di sinonimia, da parte di soggetti non-esperti. In effetti, l'assegnazione di ranghi ad un insieme di elementi (parole, concetti, oggetti, ecc.) richiede un procedura di ordinamento cui fare riferimento : a tal proposito, la letteratura psicometrica ha dato grande rilievo al criterio dei confronti appaiati (*paired comparisons*), in base al quale la formulazione di graduatorie di m elementi deriva dagli $m(m-1)/2$ confronti tra tutte le possibili coppie di *items* (Bradley e Terry, 1952). E' risultato, altresì, vero che tale procedura è accompagnata da una componente di incertezza che, generalmente, aumenta con m e caratterizza soprattutto i ranghi assegnati a quegli elementi verso i quali non esiste un giudizio netto da parte del soggetto.

Con riguardo alle graduatorie di sinonimia, possiamo assumere che il rango assegnato ad un sinonimo s_j di una parola w sia il risultato di un processo articolato in due componenti, che intervengono con ruoli distinti. La prima componente è relativa alla valutazione (di ciascun soggetto) concernente la sinonimia di s_j rispetto a w : ci sembra lecito ipotizzare che tale valutazione avvenga secondo uno schema di confronti appaiati, mediante il quale confrontando tutte le possibili coppie di sinonimi a disposizione si individua il vocabolo maggiormente prossimo a w (quello "vincente" in tutti i confronti), e così via[3]. La seconda componente, invece, esprime l'incertezza che è presente nell'assegnazione del rango, a causa della natura intrinsecamente "sfocata" dei possibili sensi di w nella percezione linguistica.

Tali due componenti possono essere rappresentate dalle variabili casuali Binomiale traslata e Uniforme discreta, rispettivamente, mediante un modello mistura definito MUB (D'Elia e Piccolo, 2003). Sia, infatti, r il rango assegnato ad un sinonimo s_j della parola w ; allora, è possibile considerare r come una realizzazione della variabile casuale $R \sim \text{MUB}(m, \pi, \xi)$ se :

$$\Pr(R = r) = \pi P_B(r) + (1-\pi) P_U(r), \qquad r = 1, 2, \ldots, m,$$

[3] Evidentemente, è anche possibile ipotizzare che tale processo si svolga in modo gerarchico, quando si è in presenza di una parola w che ammette una lista di sinonimi S_w i cui significati sono riconducibili a k concetti ($C_1, \ldots, C_h, \ldots, C_k$) ben distinti. In tal caso, quindi, i confronti appaiati avvengono a due livelli : prima tra i concetti, e successivamente tra le parole all'interno di uno stesso concetto C_h (h=1, 2, …, k).

$$P_B(r) = \binom{m-1}{r-1}(1-\xi)^{r-1}\xi^{m-r}; \qquad P_U(r) = \frac{1}{m}; \qquad r = 1,2,...,m;$$

dove :

e $\pi \in [0, 1]$; $\xi \in [0, 1]$.

Con riferimento alle graduatorie di sinonimia percepita, i parametri del modello MUB possono così essere interpretati : m è il numero dei sinonimi di w, e come tale è fisso e noto a priori. Il parametro π è inversamente legato all'incertezza presente nel processo di formulazione di una graduatoria di sinonimia : difatti, il fattore $(1-\pi)$ è una misura dell'incertezza che compete al rango assegnato a s_j.

Per quanto riguarda, invece, il significato del parametro ξ, può essere utile considerare le due seguenti espressioni :

$$\Pr(R=1) = \pi\xi^{m-1} + (1-\pi)/m; \qquad E(R) = \pi(m-1)(1/2-\xi) + (m+1)/2.$$

Da esse, infatti, emerge che – a parità di π – al crescere di ξ aumenta Pr(R = 1), cioè la probabilità che s_j sia considerato il sinonimo più vicino alla parola w, e viceversa diminuisce $E(R)$, conducendo così ad una distribuzione di probabilità con un valore atteso del rango assegnato a s_j più basso. Ne deriva che il parametro ξ può essere interpretato come una misura della forza della sinonimia percepita : in particolare, poiché per ξ = 1/2 si ottiene una distribuzione simmetrica intorno a $E(R) = (m+1)/2$, tale valore del parametro (1/2) può essere considerato come una soglia tra sinonimia sentita in modo debole e forte. In altri termini, il parametro ξ può essere altresì considerato come una *misura della scambiabilità* tra la parola w e il sinonimo s_j, così come risulta dalle evidenze empiriche.

Per quanto concerne, poi, le stime dei due parametri (π, ξ) caratterizzanti la distribuzione, esse sono derivabili con il metodo della massima verosimiglianza, la cui complessità computazionale richiede il ricorso all'algoritmo E-M (*Expectation – Maximisation*). L'efficacia di tale procedura per la stima di modelli mistura è ampiamente documentata nella letteratura statistica (McLachlan e Krishnan, 1997 ; McLachlan e Peel, 2000), ed è stata riscontrata anche per il modello MUB (D'Elia e Piccolo, 2003).

Il modello MUB può essere esteso al fine di contemplare la presenza di covariate relative ai soggetti che esprimono le graduatorie di sinonimia. In particolare, seguendo una logica analoga a quella dei Modelli Lineari Generalizzati (McCullagh e Nelder, 1989), è possibile introdurre un legame tra il vettore di variabili esplicative $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})'$ specifico di ciascun i-esimo soggetto (i = 1, 2, …, n) e i parametri π e ξ, considerati separatamente (D'Elia, 2003) oppure congiuntamente (Piccolo, 2003).

La specificazione del modello MUB, quindi, avviene mutuando la funzione logistica dei modelli logit, mediante la quale si crea una corrispondenza tra l'insieme reale e l'intervallo [0, 1] su cui sono definiti sia π che ξ :

$$(\pi \mid \mathbf{X} = \mathbf{x}_i) = \frac{1}{1+\exp(-\mathbf{x}_i\boldsymbol{\beta})}; \qquad (\xi \mid \mathbf{X} = \mathbf{x}_i) = \frac{1}{1+\exp(-\mathbf{x}_i\boldsymbol{\gamma})}.$$

Ivi $\mathbf{X}$ rappresenta la matrice del disegno di dimensioni (n x p+1), contenente i valori delle covariate degli n soggetti, mentre $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ sono i vettori dei coefficienti relativi a tali covariate.

In particolare, se si adotta la specificazione relativa al solo parametro π, si modella una situazione nella quale le caratteristiche dei soggetti che elaborano le graduatorie di sinonimia hanno effetto solo sull'incertezza, ma non sul grado di sinonimia percepita ; viceversa, se si adopera solo la specificazione relativa al parametro ξ, ciò equivale ad assumere che il grado di sinonimia avvertita dipenda dalle covariate, ma non l'incertezza nell'assegnazione del rango. Chiaramente, l'utilizzo contemporaneo di entrambe le specificazioni conduce ad un modello nel quale sia il grado di sinonimia percepita che l'incertezza nella scelta sono interpretabili attraverso le caratteristiche dei soggetti intervistati.

Anche nel caso del modello MUB esteso con la presenza delle covariate, la stima dei parametri e/o dei coefficienti delle variabili esplicative avviene mediante il ricorso all'algoritmo E-M, opportunamente modificato per tener conto delle covariate.

Le stime ottenute possono, quindi, essere adoperate per quantificare non solo la percezione della sinonimia e la connessa incertezza, ma anche per valutare l'impatto che le caratteristiche individuali hanno in tali componenti. In particolare, mediante il modello MUB con l'inclusione di covariate è possibile individuare profili distinti di soggetti (ad esempio, in base al livello di istruzione) e misurare come i diversi livelli di una o più variabili conducano a diverse graduatorie di sinonimia percepita.

3. L'indagine

L'indagine si è svolta nell'arco di un mese mediante somministrazione di un questionario a 654 soggetti non esperti ed appartenenti ad ambiti sociali e culturali sufficientemente differenziati. Esso è articolato in due sezioni : la prima è relativa alle informazioni sul soggetto e sul contesto in cui vive e opera ; la seconda parte, invece, prevede l'elaborazione di graduatorie di sinonimia.

A tal fine sono stati prescelti vocaboli appartenenti a categorie grammaticali distinte : cioè sostantivi, aggettivi, verbi. I vocaboli scelti sono : *scolaro*, *solare*, *piantare*. Per ciascun vocabolo la lista estesa dei sinonimi è stata ottenuta a partire dai cinque maggiori dizionari italiani dei sinonimi[4], e mediante consultazione del sito web :
http ://parole.virgilio.it/parole/sinonimi_e_contrari/.

Nel questionario le liste dei sinonimi di ciascun vocabolo sono state presentate in ordine alfabetico. Ad ogni soggetto partecipante all'indagine è stato chiesto di elaborare una graduatoria dei sinonimi, ordinandoli a partire da quello che lui/lei riteneva maggiormente simile (e, quindi, scambiabile) rispetto alla parola *w* di riferimento. E'stato, inoltre, chiesto di non assegnare lo stesso rango a sinonimi distinti all'interno di una stessa lista, e di non consultare vocabolari, dizionari, grammatiche, ecc.

4. Principali evidenze empiriche

La nostra proposta per un nuovo approccio all'analisi della sinonimia consente di misurare e valutare in modo oggettivo il grado di similarità percepito tra una parola ed i suoi possibili sinonimi. Inoltre, nel caso del modello con covariate, è possibile individuare profili di utenti in base ai sensi della parola che vengono da quest'ultimi privilegiati e porre, quindi, in corrispondenza le caratteristiche individuali e il grado di sinonimia assegnato.

[4] Gabrielli (1967, Loescher) ; Pittano (1987, Zanichelli) ; Quartu (1994, Rizzoli) ; De Mauro (2002, Mondadori) ; Stoppelli (2002, Garzanti).

In questo paragrafo illustriamo alcuni risultati (i principali, per motivi di spazio) emersi dall'indagine, al fine di mettere il luce le potenzialità della metodologia proposta e descritta nel paragrafo 2.

4.1. *La stima del grado di sinonimia percepita*

Nelle tabelle 1, 2 e 3 sono illustrati i risultati della applicazione del modello MUB (senza inclusione di covariate) ai tre vocaboli considerati. Per ciascun vocabolo è riportato l'elenco dei relativi sinonimi con l'indicazione delle stime $\hat{\xi}$ e $\hat{\pi}$ dei corrispondenti parametri del modello (e i rispettivi errori standard), il rango medio $\bar{r}_n$ e la varianza dei ranghi osservati. Si è ritenuto di ordinare le graduatorie in base al valore stimato del parametro ξ, poiché, come già detto nel paragrafo 2, esso può essere interpretato come una *misura di scambiabilità* tra i vocaboli e quindi fornisce una indicazione immediata del grado di sinonimia percepito tra s_j e w. Si noti inoltre che le stime del parametro ξ dipendono dal valore $\bar{r}_n$ (media dei ranghi osservati) ma in modo non lineare : pertanto, le graduatorie basate sull'uno o sull'altro criterio possono coincidere (come nel caso del vocabolo *scolaro*), ma non necessariamente. Per quanto concerne il parametro π, si è detto nel paragrafo 2 che esso è inversamente legato all'incertezza nella formulazione della graduatoria : pertanto, per ciascun sinonimo, il complemento ad uno della stima di π, (1-$\hat{\pi}$), fornisce una misura della incertezza nell' assegnazione del relativo rango a s_j rispetto a w.

Sinonimi	$\hat{\xi}$	*e.s.*	$\hat{\pi}$	*e.s.*	$\bar{r}_n$	Var(r)
Alunno	0.924	*0.005*	0.955	*0.012*	1.561	0.802
Studente	0.817	*0.007*	0.917	*0.019*	2.249	1.349
Allievo	0.712	*0.007*	0.990	*0.011*	2.745	0.823
Educando	0.376	*0.008*	0.963	*0.022*	4.720	1.409
Discepolo	0.286	*0.008*	0.945	*0.020*	5.223	1.397
Discente	0.237	*0.009*	0.824	*0.028*	5.372	1.995
Seguace	0.125	*0.006*	0.941	*0.015*	6.130	1.202

Tabella 1. Sinonimia percepita rispetto a "scolaro".

Nel caso del vocabolo *scolaro*, il sinonimo percepito come più prossimo, e quindi dotato del maggior grado di scambiabilità, è *alunno* ; si ricordi, infatti, che essendo $\xi \in [0, 1]$ il valore 0.924 sta ad indicare una grado di scambiabilità quasi perfetto. Meno prossimi, ma comunque percepiti come altamente scambiabili sono i sinonimi *studente* ed *allievo*. I rimanenti sinonimi sono invece caratterizzati da un valore stimato di ξ di molto inferiore. Quindi, si può dire che nel caso della parola *scolaro*, vi è un nucleo forte di sinonimi formato dai vocaboli *alunno*, *studente* ed *allievo* che sono legati al senso di *scolaro* come colui che apprende delle nozioni nell'ambito di un programma educativo. Il termine *educando* che si colloca al centro della graduatoria sembra invece identificare un senso separato, mentre i tre i rimanenti sinonimi che individuano un senso della parola in oggetto come colui che apprende un credo o una dottrina, sono percepiti come meno scambiabili rispetto al vocabolo *scolaro* e quindi sono sinonimi deboli nella percezione degli utilizzatori. Come si è detto in precedenza, le stime dei parametri sono legate a $\bar{r}_n$ che in tale caso fornisce una graduatoria coincidente con quella ottenuta mediante le stime di ξ ed i cui valori, se si guarda al loro campo di variazione, rispecchiano la distinzione effettuata tra sinonimi forti e deboli.

Per quanto riguarda invece l'aspetto della incertezza nella formulazione della graduatoria, i sinonimi sono caratterizzati tutti da una bassissima incertezza, che è presumibilmente riconducibile alla ridotta numerosità dei sinonimi in questione (m = 7).

Nel caso del vocabolo *solare* l'esame della graduatoria consente di identificare una sorta di percorso che conduce dal senso di *solare* inteso come capacità (figurata) di irradiare luce (*radioso*, *raggiante*, *luminoso*, *splendente*, *brillante*, *scintillante*, *sfolgorante*) a quello di *solare* come comprensibile (*lampante*, *visibile*, *evidente*, *palese*, *indiscutibile*, *indubitabile*, *innegabile*, *lapalissiano*). Si noti come il termine *chiaro*, associabile ad entrambi i sensi, occupi una sorta di posizione di passaggio nell'ambito della graduatoria giocando un ruolo di termine "di transizione".

Sinonimi	$\hat{\xi}$	*e.s.*	$\hat{\pi}$	*e.s.*	$\bar{r}_n$	Var(r)
Radioso	0.914	*0.004*	0.648	*0.024*	3.841	9.468
Raggiante	0.913	*0.004*	0.634	*0.025*	3.940	10.185
Luminoso	0.904	*0.005*	0.602	*0.026*	4.017	9.549
Splendente	0.832	*0.005*	0.774	*0.023*	4.320	7.954
Brillante	0.763	*0.006*	0.733	*0.026*	5.115	8.019
Scintillante	0.675	*0.007*	0.719	*0.028*	6.746	9.287
Sfolgorante	0.625	*0.011*	0.437	*0.034*	7.879	13.861
Chiaro	0.591	*0.008*	0.666	*0.031*	6.913	9.083
Lampante	0.440	*0.009*	0.604	*0.032*	9.125	10.107
Visibile	0.422	*0.009*	0.575	*0.033*	10.003	10.508
Evidente	0.406	*0.006*	0.799	*0.026*	9.599	7.160
Palese	0.266	*0.008*	0.608	*0.030*	11.436	10.450
Indiscutibile	0.175	*0.005*	0.823	*0.021*	12.783	6.467
Indubitabile	0.146	*0.004*	0.884	*0.017*	13.318	5.416
Innegabile	0.123	*0.004*	0.825	*0.020*	13.416	6.971
Lapalissiano	0.023	*0.003*	0.395	*0.024*	13.182	9.366

Tabella 2. Sinonimia percepita rispetto a "solare".

Un ulteriore aspetto da sottolineare riguarda la incertezza nella formulazione dei ranghi che appare maggiore rispetto al caso del vocabolo *scolaro* precedentemente esaminato. I più elevati valori di (1-$\hat{\pi}$) sono da ascrivere alla maggiore lunghezza della lista di sinonimi che tende ad accrescere l'incertezza nella assegnazione del rango e anche la variabilità dei ranghi osservati. E' opportuno notare che, nonostante i differenti valori di m (7 e 16), una comparazione tra i vocaboli, in termini di incertezza, è comunque possibile, anche se non è di immediato interesse per i fini di questo lavoro.

L'ultimo vocabolo considerato è quello del verbo *piantare* che presenta una lista di sinonimi piuttosto lunga (m = 20) nell'ambito della quale è possibile individuare vari sensi.

Sinonimi	$\hat{\xi}$	*e.s.*	$\hat{\pi}$	*e.s.*	$\bar{r}_n$	Var(*r*)
Seminare	0.975	*0.003*	0.487	*0.023*	5.630	31.646
Coltivare	0.949	*0.004*	0.467	*0.025*	6.219	32.755
Interrare	0.932	*0.005*	0.350	*0.026*	6.844	28.743
Innestare	0.860	*0.013*	0.203	*0.028*	9.257	35.291
Conficcare	0.744	*0.018*	0.191	*0.031*	9.183	27.064
Ficcare	0.721	*0.013*	0.279	*0.032*	9.353	23.571
Infilare	0.658	*0.013*	0.327	*0.033*	9.671	21.900
Inserire	0.616	*0.012*	0.348	*0.033*	9.587	20.105
Introdurre	0.601	*0.015*	0.281	*0.033*	10.783	24.347
Mettere	0.552	*0.015*	0.281	*0.034*	10.641	24.447
Collocare	0.511	*0.015*	0.280	*0.034*	10.392	24.515
Porre	0.479	*0.013*	0.348	*0.034*	11.477	21.720
Sistemare	0.428	*0.016*	0.268	*0.023*	12.838	23.643
Mollare	0.216	*0.012*	0.273	*0.030*	12.275	32.019
Abbandonare	0.214	*0.026*	0.118	*0.029*	11.618	40.371
Lasciare	0.207	*0.011*	0.293	*0.030*	11.963	33.133
Cessare	0.174	*0.011*	0.267	*0.029*	12.635	31.376
Smettere	0.170	*0.010*	0.318	*0.029*	12.446	33.645
Interrompere	0.155	*0.007*	0.423	*0.029*	13.765	26.608
Troncare	0.038	*0.008*	0.154	*0.022*	12.593	35.330

Tabella 3. Sinonimia percepita rispetto a "piantare".

Innanzitutto, si osservi che tale verbo viene percepito come avente un senso legato alla attività agricola con tre sinonimi (*seminare*, *coltivare* ed *interrare*) caratterizzati da un grado di scambiabilità molto elevato e un quarto (*innestare*) meno marcato, ma comunque percepito come altamente scambiabile. Un secondo senso è individuato dai sinonimi *conficcare*, *ficcare*, *infilare*, *inserire*, *introdurre*, caratterizzati da un grado percepito di scambiabilità medio. Gli ultimi due sensi individuabili sono costituiti da sinonimi deboli : *piantare* come collocare in un posto o ordine (*mettere*, *collocare*, *porre*, *sistemare*) e *piantare* come l'atto di porre fine. Si noti come in questo caso, per tutti i sinonimi il grado di incertezza sia particolarmente elevato.

4.2. *L'effetto delle caratteristiche individuali sulla percezione della sinonimia*

Il modello MUB per l'analisi delle graduatorie di sinonimia consente anche di includere nella sua specificazione la presenza di covariate, vale a dire di variabili esplicative relative alle caratteristiche individuali, mediante le quali è possibile interpretare le graduatorie espresse.

In particolare, sembra interessante poter individuare quali variabili condizionano la posizione in graduatoria di un fissato sinonimo s_j (j=1, 2,…, m) rispetto alla parola obiettivo w, e quantificare direzione e forza di tale impatto. Inoltre, l'opportuna combinazione di valori delle variabili esplicative, risultate rilevanti, permette la definizione di profili di utenti della

lingua, la cui percezione di sinonimia del termine s_j rispetto a *w* appare significativamente differenziata.

Al fine di evidenziare tali potenzialità, illustriamo di seguito solo alcuni risultati emersi dall'indagine condotta rispetto alle parole obiettivo, le cui stime delle graduatorie di sinonimia percepita sono state commentate nel precedente paragrafo 4.1.

• Con riferimento al sostantivo *scolaro*, discutiamo le evidenze emerse per il sinonimo *discente* (che presenta un grado di scambiabilità molto modesto, cioè $\hat{\xi}$ = 0.237). Per tale sinonimo sono risultate significative rispetto al parametro ξ (misura di sinonimia) le variabili esplicative "numero di componenti della famiglia", "possesso della laurea", "lettore assiduo (di libri)", come si evince dalla Tabella 4. La misura dell'incertezza, espressa da $(1 - \hat{\pi})$, invece non è risultata dipendere da alcuna covariata.

Covariate	**Stime ($\hat{\gamma}$)**	**Errori standard**
Costante	-0.831	*0.167*
Numero componenti famiglia	-0.115	*0.038*
Laurea (NO=0, SI=1)	0.896	*0.110*
Lettore assiduo (NO=0,SI=1)	0.290	*0.121*
	($\hat{\pi}$)	**Errore standard**
	0.867	*0.013*

Tabella 4. Modello MUB con covariate per il sinonimo "discente".

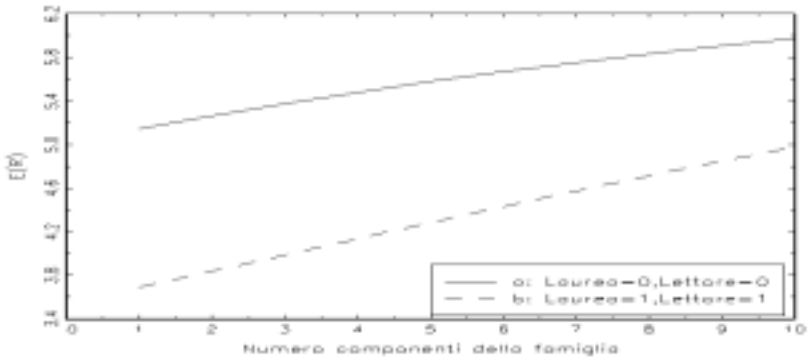
Le stime ottenute evidenziano che il possesso della laurea e il fatto di essere lettori assidui esercitano un impatto positivo sul grado di sinonimia percepita, in quanto determinano un aumento del parametro ξ, che può essere considerato come una misura di scambiabilità del sinonimo *discente* rispetto alla parola obiettivo *scolaro*. Un impatto di segno opposto è, invece, esercitato dalla variabile "numero di componenti della famiglia".

Nella successiva Tabella 5 sono riportati, per alcuni possibili profili di parlanti, i corrispondenti valori attesi del rango assegnato a *discente* nella graduatoria di sinonimia e le corrispondenti stime del parametro ξ.

Numero componenti famiglia	Laurea (NO=0, SI=1)	Lettore assiduo (NO=0, SI=1)	$\hat{\xi}$	E(*R*)
3	0	0	0.236	5.376
6	0	0	0.180	5.670
3	1	1	0.503	3.985
6	1	1	0.418	4.429

Tabella 5. Profili e valori attesi del rango per il sinonimo "discente"

In tal modo, emerge che un profilo culturale elevato (laureato e lettore assiduo) insieme con l'appartenenza ad una famiglia poco numerosa determinano il più alto grado ($\hat{\xi}$ = 0.503) di sinonimia percepita della parola *discente* rispetto a *scolaro* ; all'opposto si colloca il profilo



culturale modesto congiunto alla provenienza da famiglie abbastanza numerose ($\hat{\xi}$ = 0.180). Ciò appare giustificabile, se si considera che nella lingua italiana il termine *discente* non risulta essere di uso comune, ed è da considerarsi una parola cosiddetta "colta", per la quale è lecito ipotizzare che esistano difficoltà di attribuzione di senso da parte di persone di media/modesta cultura.

Il ruolo svolto dal "numero di componenti della famiglia" in rapporto al valore atteso del rango assegnato a *discente* è evidenziato anche nella Figura 1, che conferma come al crescere del "numero di componenti" diminuisca la percezione di sinonimia rispetto a *scolaro* (in quanto aumenta il rango atteso).

• Per quanto riguarda il verbo *piantare*, illustriamo qui i risultati ottenuti per il sinonimo *seminare*, che presenta il maggior grado di scambiabilità stimato ($\hat{\xi}$ = 0.975), come si evinceva dalla precedente Tabella 3.

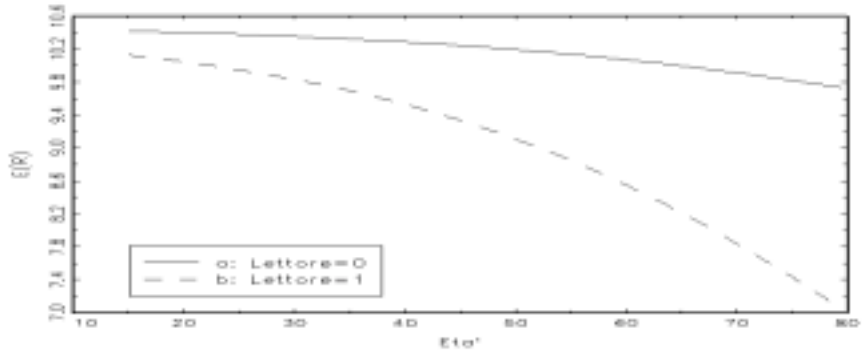
Per tale sinonimo sono risultate significative rispetto al parametro ξ (misura di sinonimia) la variabile esplicativa "lettore assiduo", e rispetto al parametro π la variabile "età" (Tabella 6).

Covariate	**Stime ($\hat{\gamma}$)**	**Errori standard**
Costante	0.446	*0.088*
Lettore assiduo (NO=0,SI=1)	5.049	*0.797*
	($\hat{\beta}$)	**Errori standard**
Costante	-3.798	*0.380*
Età	0.041	*0.010*

Tabella 6. Modello MUB con covariate per il sinonimo "seminare".

Le stime ottenute evidenziano, anche in questo caso, che il fatto di essere lettori assidui esercita un impatto positivo sul grado di sinonimia percepita, in quanto determina un aumento del parametro ξ, che può essere considerato come una misura di scambiabilità del sinonimo *seminare* rispetto alla parola obiettivo *piantare*. Per quanto concerne, invece, la variabile "età", essa influisce sul grado di incertezza : infatti, il segno positivo della stima del rispettivo coefficiente evidenzia che all'aumentare dell'"età" cresce anche $\hat{\pi}$, diminuendo quindi l'incertezza.

In effetti, poiché il valore atteso del rango di un sinonimo s_j rispetto a w dipende sia dal grado di scambiabilità percepita che dalla misura di incertezza, entrambe le variabili individuate ("lettore assiduo" ed "età") esercitano un ruolo nel determinare la posizione attesa di *semi-*



nare nella graduatoria, condizionatamente a determinati profili di utenti. Ciò è evidenziato nella successiva Figura 2.

Dalla Figura emerge che la variabile "lettore assiduo" diventa via via più rilevante al crescere dell'"età", con la quale, evidentemente, interagisce. In particolare, appare che il senso di *piantare* inteso come *seminare* è privilegiato da persone più adulte e lettori assidui, che quindi prediligono l'accezione più tradizionale del verbo in questione.

5. Considerazioni finali

L'articolo ha evidenziato come nello studio della sinonimia percepita possa essere utile il ricorso a modelli per variabili rango, che permettono da un lato di stimare in modo oggettivo sia il grado di sinonimia che l'eventuale misura di incertezza presente nel processo, e d'altro canto consentono l'individuazione e la quantificazione dell'impatto esercitato su tale percezione dalle caratteristiche personali dei parlanti. Tali potenzialità possono essere utilmente sfruttate in numerosi ambiti, nei quali è importante associare ad un termine il senso privilegiato dalla maggior parte degli utenti, o da una parte di essa con particolari caratteristiche (si pensi, ad esempio, al problema della cosiddetta *information retrivial*).

Ulteriori sviluppi sono possibili sia dal punto di vista metodologico che in termini di campi di applicazione. Sul primo versante, infatti, sarebbe utile specificare un modello per ranghi di tipo gerarchico, in modo da tener conto del fatto che l'elaborazione di una graduatoria di sinonimia può avvenire mediante due stadi : graduatoria tra i differenti sensi e, poi, assegnazione dei ranghi all'interno di un singolo senso. Dal punto di vista delle applicazioni, invece, sarebbe interessante investigare il ruolo della sinonimia percepita con riferimento a lingue diverse, per cogliere l'eventuale presenza di strutture semantiche analoghe.

Ringraziamenti : Il presente lavoro è stato svolto nell'ambito dei progetti di ricerca afferenti al Dipartimento di Scienze Statistiche, Università di Napoli Federico II ; ci si è inoltre avvalsi dei fondi della L.R. 5/2002.

Bibliografia

Bolasco S. (1999). *Analisi Multidimensionale dei Dati*. Carocci.

Bradley R.A. e Terry M.A. (1952). Rank analysis of incomplete block designs I. *Biometrika*, vol. (39) : 324-345.

Cappelli C. (2003). Identifying word senses from synonyms : a cluster analysis approach. *Quaderni di Statistica*, vol. (5) : 105-117.

Cappelli C. e Corduas M. (2003). Assessing synonymy links : a cluster analysis approach. In *Book of Short Papers CLADAG 2003* : 91-94.

D'Elia A. (2003). A mixture model with covariates for ranks data : some inferential developments. *Quaderni di Statistica*, vol. (5) : 1-25.

D'Elia A. e Piccolo D. (2002). Analisi statistica delle preferenze : metodi e modelli a confronto. In Frosoni B., Magagnali U. e Boari G. (Eds), *Studi in onore di Angelo Zanella*. Vita e Pensiero : 167-187.

D'Elia A. e Piccolo D. (2003). A mixture model for preferences data analysis. *Submitted*.

De Mauro T. (2002). *Dizionario della lingua italiana*. Mondadori.

Lebart L., Salem A. e Berry L. (1998). *Exploring Textual Data*. Kluwer Academic Publishers.

Lin D., Zhao S., Qin L. e Zhou M. (2003). Identifying Synonyms among Distributionally Similar Words. In *Proceedings of IJCAI-03* : 1492-1493.

Marden J.I. (1995). *Analyzing and Modeling Rank Data*. Chapman & Hall.

McCullagh P. e Nelder J. (1989). *Generalized Linear Models (2nd edition)*. Chapman & Hall.

McLachlan G. e Krishnan T. (1997). *The E-M Algorithm and Extensions*. J. Wiley & Sons.

McLachlan G. e Peel D. (2000). *Finite Mixture Models*. J. Wiley & Sons.

Piccolo D. (2003). Computational issues in the E-M algorithm for ranks model estimation with covariates. *Quaderni di Statistica*, vol. (5) : 27-48.

Ploux S. e Ji H. (2003). A model for matching semantic maps between languages (French/English, Englissh/French). *Computational Linguistics*, vol. (29) : 155-178.

Woods A., Fletcher P. e Hughes A. (1986). *Statistics in Language Studies*. Cambridge University Press.

Gli aggettivi delle rappresentazioni di genere in adolescenza

Simona Carbone, Maria Longobardi

IRMC – Istituto di Ricerca sulla Multimedialità e la Comunicazione della Provincia di Caserta
Via Santa Chiara – Palazzo Mirabella – 81100 Caserta – Italia
simona.carbone@tin.it, streghemjd@libero.it

Riassunto

L'adolescenza è il momento determinante nella vita di un individuo per la formazione della sua identità; le definizioni in relazione al genere che un individuo elabora in questo periodo sono fondamentali per la determinazione della sua personalità.
Attraverso un questionario sono state rilevate le rappresentazioni di un gruppo di adolescenti di un coetaneo dello stesso sesso e di quello opposto, quindi le rappresentazioni che gli adolescenti hanno del proprio genere e di quello opposto.
Con l'analisi lessicale degli aggettivi utilizzati è stato possibile descrivere queste rappresentazioni, e, attraverso la ricerca delle forme specifiche, in positivo e in negativo, si è inteso cogliere le peculiarità delle rappresentazioni che maschi e femmine hanno dell'identità di genere propria e opposta.
Infine, con l'analisi delle corrispondenze lessicali, si sono descritte le caratteristiche delle risposte rispetto alle categorie di contenuto.
Ne è risultata una maggiore capacità interpersonale delle ragazze rispetto ai ragazzi, maggiormente interessati, questi ultimi, alla dimensione estetico-corporea.

Abstract

Adolescence is a basic moment in each person's life for the building up of its own identity; all the definitions regarding the genre that an individual can work out in this period are essential to determine his own personality.
Through a questionnaire it was possible to point out the descriptions of a group of teen-agers, of a contemporary person from the same sex and another one from the opposite sex, thus the descriptions of himself and of another person.
Through a lexical analysis of the adjectives here used, it was possible to describe these representations and, through the search of specific forms, both positive and negative, we meant to examine the peculiarity of the representations that males and females can have on the identity of genre -both of their own identity and of the opposite side. Finally, through an analysis of lexical correspondences, it was possible to describe the characteristics of each answer compared to some categories of contents.
What clearly emerges from this analysis is that girl are endowed of a greater interpersonal capability than boys.
The latters seem more interested to the aesthetical-physical dimension.

Parole chiave: rappresentazioni, rappresentazioni di genere, aggettivi, comparazione tra sub-testi, forme caratteristiche, corrispondenze lessicali.

1. Introduzione

Il genere è lo sfondo su cui ciascuno rappresenta la propria vita, la variabile che più di ogni altra caratterizza ogni individuo: prendendo in esame gli aspetti più minuti della vita quotidiana, non ve n'è alcuno che non sia connotato secondo il genere (Burr, 2000).

Il nostro interesse verso le rappresentazioni di genere è derivato dalla consapevolezza dei profondi cambiamenti avvenuti nel rapporto uomo/donna nella società contemporanea, caratterizzata, oggi, dalla crisi dei ruoli legati ai generi.

Specifiche qualità sembrano essere attribuite ai soggetti, maschili e femminili, in virtù del

sistema di pensiero e di norme di riferimento di chi giudica ed è interessante riflettere, a nostro avviso, sul loro grado di influenza sui pensieri, sentimenti e azioni.

Con questa indagine si è tentato di focalizzare l'attenzione sulla ***rappresentazione di genere*** e sulla sua ***interiorizzazione***; concetto che sembra offrire, nella ricerca più recente, migliori occasioni per comprendere il complesso rapporto tra i sistemi cognitivi complessi presenti in ogni individuo e i sistemi di rapporti simbolici esistenti tra gli attori sociali (Palmonari, 1989).

Il concetto di ***rappresentazione*** elaborato dallo studioso Moscovici è definito come "un sistema di valori, idee, prassi" che svolge la duplice funzione di stabilire un ordine che permetta agli individui di orientarsi nel loro mondo sociale e materiale e rendere possibile la "comunicazione tra i membri di una comunità fornendo loro un codice per gli scambi sociali" (Moscovici, 1984). La rappresentazione sociale è sia un meccanismo psichico, in quanto espressione della mente umana, che un meccanismo sociale, in quanto prodotto culturale. Le scelte personali di un soggetto possono variare in relazione alla rappresentazione di sé e della realtà in cui vive.

2. Le differenze di genere

La distinzione fra differenze di genere e differenze sessuali è di fondamentale importanza, perché questi due costrutti rimandano a due diversi presupposti. Si fa riferimento alle ***differenze di sesso*** per la distinzione essenzialmente biologica che si fonda sulle caratteristiche anatomiche e fisiologiche degli individui, mentre si rimanda al concetto di ***genere*** in una prospettiva sociale, per sottolineare una caratteristica socioculturale che descrive comportamenti e stili riconosciuti propri di ciascun gruppo. Le caratteristiche di genere si riferiscono al *significato sociale* assunto dalle differenze sessuali, ai comportamenti che vengono associati ai maschi e alle femmine e, di conseguenza, attesi all'interno di un particolare contesto sociale. E' evidente la tendenza crescente nelle scienze sociali ad insistere sulla necessità di tale distinzione.

Maschile e femminile sono categorie astratte, costrutti che variano a seconda dei contesti, delle regole condivise e delle interiorizzazioni sociali. Non esiste un'unica modalità femminile o maschile di esprimersi, ma tanti modi di essere maschio e femmina.

Il vantaggio di considerare la distinzione tra l'appartenenza ad un gruppo sessuale e le identità di genere sta nel riuscire a considerare le variazioni dei comportamenti definiti secondo il genere all'interno dei gruppi sessuali. Ciò ci permette di sfuggire all'assunzione che tutti i membri di un gruppo sessuale adottano i ruoli di genere nella stessa maniera.

"Il tema delle differenze di genere non contempla questioni scontate. Non si trovano rappresentazioni del tutto consolidate del rapporto fra uomini e donne nella opinione pubblica, né tanto meno nelle discipline scientifiche che possono offrire un contributo alla comprensione dei problemi che assumono lo statuto di fenomeni sociali" (Burr, 2000).

A partire dagli anni Settanta, le definizioni di maschile e femminile sono divenute sempre più complesse e differenziate. Lo schema bipolare che opponeva maschile e femminile non era più considerato idoneo per misurare il comportamento legato al ruolo sessuale individuale.

Si è cominciato a sottolineare il carattere multidimensionale della socializzazione dei ruoli sessuali (Huston, 1985) e l'importanza delle aspettative verso il ruolo; importante è diventato considerare in primo piano l'influenza delle aspettative legate al ruolo sessuale.

In merito alle differenti performance, Halpern e Jones hanno mostrato che le prestazioni di maschi e femmine sono modificabili grazie all'esercizio e alle differenti aspettative. Ricon-

siderare le differenze tra i sessi come altamente sensibili alle modificazioni dell'ambiente ha rimesso ancora più profondamente in discussione la psicologia delle differenze di genere.

In una prospettiva psicosociale, Eagly (1987) ha affrontato la psicologia delle differenze di genere, considerando i problemi metodologici legati a questo tipo di studi. In particolare, ha messo in evidenza il contrasto esistente tra due filoni di ricerca, gli studi riguardanti le differenze di genere e le indagini condotte tra il grande pubblico, che hanno evidenziato le diverse convinzioni che la gente possiede rispetto alle differenze di genere.

Gli studi relativi agli stereotipi legati di sessi indicano che, nella vita sociale, le donne sono percepite come continuamente interessate al benessere altrui, mentre gli uomini sono molto più interessati a sé stessi e alla direttività. L'analisi della Eagly, sorretta da un modello concettuale chiaramente formulato e corredata da dati empirici, ha dato validità agli stereotipi riguardanti il comportamento sociale dei sessi (Lloyd, 1994). La studiosa ha messo a punto un modello concettuale basandosi sulla teoria dei ruoli e sulle ricerche effettuate nel campo dell' influenza sociale, anziché porre l'accento sui processi cognitivi, come fanno i sociopsicologi contemporanei nelle loro procedure (Lloyd, 1994). Secondo la Eagly è da tenere in forte considerazione l'appartenenza a un gruppo e le pressioni sociali esercitate sull'individuo in quanto appartenente ad un sesso. Appartenendo a gruppi di sesso diverso, uomini e donne hanno aspettative diverse rispetto al ruolo e alle esperienze professionali, e di conseguenza hanno competenze e convinzioni differenti riguardanti il comportamento.

Il ruolo attribuito a ciascun sesso è rappresentato da un insieme di aspettative consensuali che non riguardano solo il comportamento dell'individuo, ma anche quello degli altri.

Attualmente il concetto di "identità di genere" è ampiamente utilizzato nell'ambito della ricerca psicologica, che non rinuncia a collocare la soggettività umana in un contesto storico culturale dato. Esso rimanda al meccanismo di interrelazione tra individuale e sociale che evidenzia il grado di influenza di questo ultimo nella strutturazione e nello sviluppo della personalità umana.

Nell'essere maschio o femmina c'è una differenza biologica ma anche una differenza legata al "io cosa mi sento", al "cosa le persone intorno a me ritengono che io sia". L'identità di genere ha una componente: biologica (come sono fatto), soggettiva (come io mi sento), ma anche sociale (come gli altri mi vedono), cognitiva (come io mi vedo), educativa (come sono stato cresciuto), culturale (quale dimensione il maschile-femminile hanno nel contesto di vita in cui vivo). E' meglio definire il costrutto di Stoller (1968) e riferirci a un costrutto multidimensionale (Arcidiacono, 1994), che la letteratura anglosassone esprime nel più generale concetto di Gender Identities.

Abbiamo focalizzato la nostra attenzione sull'adolescenza perché si configura come il momento più importante per la conferma della identità di genere, come è descritto nella letteratura psicodinamica (Nunziante Cesàro, 1998). Infatti, questa è una delicata fase di transizione nello sviluppo psico-fisico dell'individuo, caratterizzata da indeterminatezza, da cambiamenti biologici e sociali, ci si riscopre a pensarsi come non si era mai fatto, e importanti sono le definizioni in relazione al genere..

Lo scopo di questa ricerca è, dunque, lo studio delle rappresentazioni che gli adolescenti hanno dei coetanei del proprio genere e dell'altro, delle idealizzazioni e delle aspettative legate a tali rappresentazioni, anche influenzati dagli stereotipi e dai pregiudizi interiorizzati dal soggetto.

3. Metodologia della ricerca

I soggetti dell'indagine sono un campione di 179 adolescenti, di età compresa tra i 13 e i 18 anni, di cui 89 maschi e 90 femmine, appartenenti a famiglie di istruzione e livello economico medi.

Lo strumento di rilevazione utilizzato è un questionario, costruito *ad hoc* per l'esplorazione delle rappresentazioni. Agli intervistati è stato chiesto di descrivere un coetaneo di sesso opposto e uno del proprio sesso, due ragazzi "immaginari", chiamati convenzionalmente *Andrea* e *Sofia*.

Il questionario è stato elaborato in duplice versione, differente per i maschi e per le femmine. Ai maschi è stato chiesto di descrivere due ragazzi, Sofia e Andrea, di 15 anni; alle femmine le domande sono state poste in ordine inverso: descrivere Andrea e Sofia.

La prima domanda è volta a rilevare le rappresentazioni che gli adolescenti hanno dell'altro genere, la seconda domanda è volta ad indagare le rappresentazioni che gli adolescenti hanno del proprio genere.

Le risposte al questionario sono state analizzate attraverso l'utilizzo di metodi statistici di analisi dei dati testuali, con lo scopo di tentare di acquisire informazioni sulla costruzione delle rappresentazioni di genere degli adolescenti.

Per illustrare le caratteristiche delle rappresentazioni di sé e del sesso opposto e sottolineare le differenze e le peculiarità per genere, il corpus delle risposte è stato suddiviso in 4 sub-testi ottenuti incrociando il sesso dell'intervistato e quello del soggetto da descrivere, che sono stati analizzati in maniera comparativa.

In particolare, le differenze tra maschi e femmine sono state rilevate dalla ricerca delle specificità, in senso positivo o negativo, cioè delle parole specifiche di ogni sub-testo o assenti in essi, e delle parole comuni a tutti.

Con l'analisi delle corrispondenze lessicali, che confronta i diversi profili lessicali dei sub-testi, esplorando la somiglianza tra essi, con metodi di tipo fattoriale, sono stati cercati i fattori principali, le dimensioni di senso contenenti informazioni sul contesto del discorso.

4. Analisi dei dati

4.1. *Caratteristiche del corpus*

L'analisi è stata effettuata sugli aggettivi utilizzati dagli intervistati, in quanto, quali unità di contenuto del discorso, che cioè definiscono attributi e qualità, sono stati considerati le uniche parole utili a rappresentare le definizioni di caratteristiche, che è lo scopo dello studio.

Gli aggettivi sono stati riportati tutti al lemma, per ottenere la massima rappresentatività di ciascuno dalla fusione delle frequenze delle varie flessioni. Le informazioni sul genere degli aggettivi non si sono comunque perse, in quanto l'analisi è stata condotta distinguendo anche in base al soggetto da descrivere.

Gli aggettivi presenti nel corpus sono 276, per un totale di 1.897 occorrenze. La ricchezza lessicale degli aggettivi è pari a 14,5%, mentre il numero degli hapax è 107 (Tabella 1).

Da un primo sguardo alle caratteristiche dei sub-testi, emerge che le femmine hanno un lessico più ricco dei maschi; infatti, hanno usato nelle loro descrizioni un numero di aggettivi differenti superiore ai maschi.

SUB-TESTI	OCC.	FORME GRAFICHE	HAPAX	RICCHEZZA LESSICALE (FG/OCC)
Femmine-Sofia	587	133	60	7,1%
Femmine-Andrea	461	154	86	8,1%
Maschi-Sofia	468	100	43	5,3%
Maschi-Andrea	382	121	61	6,4%
TOTALE	1.897	276	107	14,5

Tabella 1. Caratteristiche del corpus e dei sub-testi

Un risultato più interessante, però, è che sia le femmine sia i maschi hanno descritto con maggiore varietà il ragazzo, Andrea, rispetto alla ragazza, Sofia. Probabilmente le femmine sono più interessate all'altro sesso, che immaginano e descrivono con molta ricchezza, mentre i ragazzi sono più narcisisti, più propensi a descrivere sé stessi.

4.2. *Analisi lessicale*

Con l'analisi delle frequenze lessicali si è inteso ottenere una prima descrizione delle rappresentazioni degli adolescenti, relativo agli aggettivi utilizzati più di frequente dai soggetti per descrivere Sofia e Andrea.

Gli aggettivi utilizzati con maggiore frequenza sono "simpatico" ed "intelligente", che, insieme, costituiscono il 17,2% del corpus; seguono "bello", "carino", "alto", "dolce", tutti indicanti caratteristiche positive (Tabella 2).

4.3. *Forme caratteristiche e forme banali*

Dopo l'analisi del corpus nel suo insieme sono stati confrontati tra loro i sub-testi di cui è composto, in modo da cogliere le differenze in base al sesso, analizzando le 'parole chiave' di ciascuno, e le caratteristiche comuni, che emergono dalle forme usate in maniera diffusa.

FORME	FREQ. TOT	Freq. F-Sofia	Freq. F-Andrea	Freq. M-Sofia	Freq. M-Andrea
Simpatico	218	72	44	60	42
Intelligente	117	34	25	35	23
Bello	90	16	25	45	4
Carino	86	31	17	31	7
Alto	82	12	15	32	23
Dolce	65	35	19	11	0
Socievole	45	13	10	7	15
Sincero	43	23	8	6	6
Estroverso	38	19	9	4	6
Disponibile	33	13	6	8	6
Allegro	31	17	9	4	1
Gentile	25	10	5	8	2
Divertente	24	9	8	2	5
Amichevole	20	5	4	1	10
Sensibile	20	9	7	2	2

Tabella 2. Forme più frequenti

Le forme banali, usate diffusamente da maschi e femmine, sono, nelle alte frequenze, gli aggettivi "simpatico" e "intelligente": per descrivere Sofia "simpatico" è stato usato 60 volte dai maschi e 72 dalle femmine, "intelligente" 35 volte dai maschi e 34 dalle femmine, per descrivere Andrea "simpatico" ha frequenza 42 per i maschi e 44 per le femmine, "intelligente" 23 per i maschi e 25 per le femmine) (Tabella 2).

Passando all'analisi delle forme caratteristiche in ogni sub-testo per sovra e sotto utilizzo (Tabella 3), le femmine descrivono Sofia rispetto alla relazione amicale, come una ragazza "comprensiva", "dolce", "allegra", "sincera", usando meno della media aggettivi relativi all' aspetto fisico, quali "alto", "bello", "attraente". Anche Andrea è descritto rispetto al carattere, "spiritoso", "profondo", ma anche "immaturo" e "scontroso". Le femmine, dunque, descrivono un loro coetaneo rispetto a caratteristiche relazionali, relative al comportamento, mentre sono poco attente al suo aspetto fisico.

I maschi, invece, rappresentano una ragazza per l'aspetto fisico, con precisi riferimenti erotici: "formoso", "sexy", non tenendo conto dell'aspetto relazionale: "comprensivo", "amicale", "altruista" sono aggettivi sotto utilizzati. Particolare è la descrizione che danno di un ragazzo: gli attribuiscono caratteri di spiacevolezza fisica, definendolo "basso", "brutto", "grasso", sotto-utilizzando aggettivi relativi al comportamento: "dolce", "affettuoso", "comprensivo".

FEMMINE-SOFIA			FEMMINE-ANDREA			MASCHI-SOFIA			MASCHI-ANDREA		
FORMA	f	F	FORMA	f	F	FORMA	f	F	FORMA	f	F
SPECIFICITA' POSITIVE											
Comprensivo	18	14	Spiritoso	13	7	Bello	90	45	Basso	15	12
Dolce	65	35	Immaturo	6	4	Formoso	10	10	Brutto	11	8
Allegro	31	17	Forte	4	3	Attraente	13	9	Grasso	10	7
Sincero	43	23	Profondo	5	4	Sexy	6	5	Pigro	6	5
			Scontroso	3	3				Antipatico	12	7
SPECIFICITA' NEGATIVE											
Affascinante	18	2	Basso	15	0	Altruista	11	0	Comprensivo	18	0
Attraente	13	0	Brutto	11	0	Spiritoso	13	0	Bello	90	4
Bello	90	16	Bravo	20	1	Amichevole	20	1	Dolce	65	0
Alto	82	12				Comprensivo	18	0			

Tabella 3. Forme specifiche nei sub-testi

In sintesi, per Andrea e Sofia viene data una duplice descrizione, basata, per le femmine, sulla personalità e sulle capacità nei rapporti interpersonali, e, per i maschi, sulla dimensione corporea, che assume connotazioni di spiacevolezza quando questi descrivono un coetaneo dello stesso sesso.

Le femmine danno di Andrea una descrizione basata sul comportamento, che risulta essere un ragazzo ben diverso dal classico *macho*. Ciò conferma la tendenza, tipica di una recente trasformazione sociale e culturale, in cui gli stereotipi dell'uomo "rude" e della donna subalterna e remissiva sembrano tramontati (Breen, 1998).

I maschi, invece, descrivono Andrea limitandosi al suo aspetto fisico, e gli attribuiscono caratteristiche negative. In generale i maschi, rispetto alle femmine, interiorizzano meno precocemente le caratteristiche di genere, e quindi risultano essere più immaturi; l'uso di aggettivi negativi potrebbe essere l'espressione di un minor livello di autostima, dovuto ad un'identità di genere ancora incerta.

Sofia è descritta dalle femmine per le sue doti di buona amica, mentre i maschi, sono molto più interessati alle sue caratteristiche fisiche, che esagerano in maniera evidente.

Le femmine, dunque, sono competenti nella sfera relazionale, mentre i maschi vengono colpiti essenzialmente dall'aspetto fisico, confermando la loro scarsa propensione all' espressione del mondo emotivo.

4.4. *Analisi delle corrispondenze lessicali*

Con questa analisi è stato possibile descrivere e sintetizzare le caratterizzazioni delle risposte rispetto alle categorie di contenuto.

I primi due assi spiegano il 72% della varianza totale, il primo asse da solo il 40,9%. L'analisi del primo piano fattoriale è pertanto più che soddisfacente.

In base ai i valori delle coordinate e dei contributi sui primi due assi fattoriali, sono state individuate due dimensioni caratterizzanti le interviste in base ai loro contenuti.

Il primo asse contrappone le femmine ai maschi, e gli aggettivi con connotazione positiva a quelli con connotazione negativa. Questa dimensione si riferisce, pertanto, alle *caratteristiche delle rappresentazioni distinte per genere*, che sono positive per le femmine e negative per i maschi.

Il secondo asse contrappone gli aggettivi che si riferiscono all'aspetto fisico agli aggettivi che definiscono comportamenti, per cui definisce il *tema oggetto della descrizione*.

Sul primo quadrante del grafico sono proiettati gli aggettivi utilizzati dai maschi per descrivere Sofia: "appariscente", "attraente", "bello", "stupendo", "formoso", "sexy", aggettivi che si riferiscono all'aspetto fisico.

Sul secondo quadrante sono proiettati gli aggettivi utilizzati delle femmine, che hanno descritto per Andrea e Sofia riferendosi ai comportamenti e alle relazioni, in chiave positiva: "sensibile", "sincero", "interessante", "affidabile".

Il terzo quadrante comprende le rappresentazioni che i maschi hanno di un coetaneo, espresse con gli aggettivi: "brutto", "grasso", "insignificante", "stupido", "egoista", ma anche "amichevole", "socievole". Le valutazioni, dunque, sono relative all'aspetto fisico e comportamentale, ed espresse in prevalenza in chiave denigratoria.

Mentre le ragazze descrivono l'aspetto comportamentale ed emotivo di un soggetto, i ragazzi sono più attenti all'aspetto fisico, in particolare alle connotazioni sessuali femminili .

Questi risultati costituiscono essenzialmente una conferma 'grafica' dei risultati ottenuti con l'analisi delle specificità.

5. Conclusioni

Per Sofia viene data una duplice descrizione basata sulla corporeità (dai maschi) e sulla personalità, con valorizzazione delle capacità interpersonali (dalle femmine), riproponendo il consueto stereotipo, che presenta le femmine competenti nelle capacità relazionali. Anche per quanto riguarda Andrea sembra emergere che i maschi siano meno orientati all'espressione del proprio mondo emotivo.

Si può quindi ipotizzare che in adolescenza le rappresentazioni di maschi e femmine, intese rispettivamente come scarsa capacità relazionale e orientamento all'altro, siano la rappresentazione di genere maggiormente rilevante.

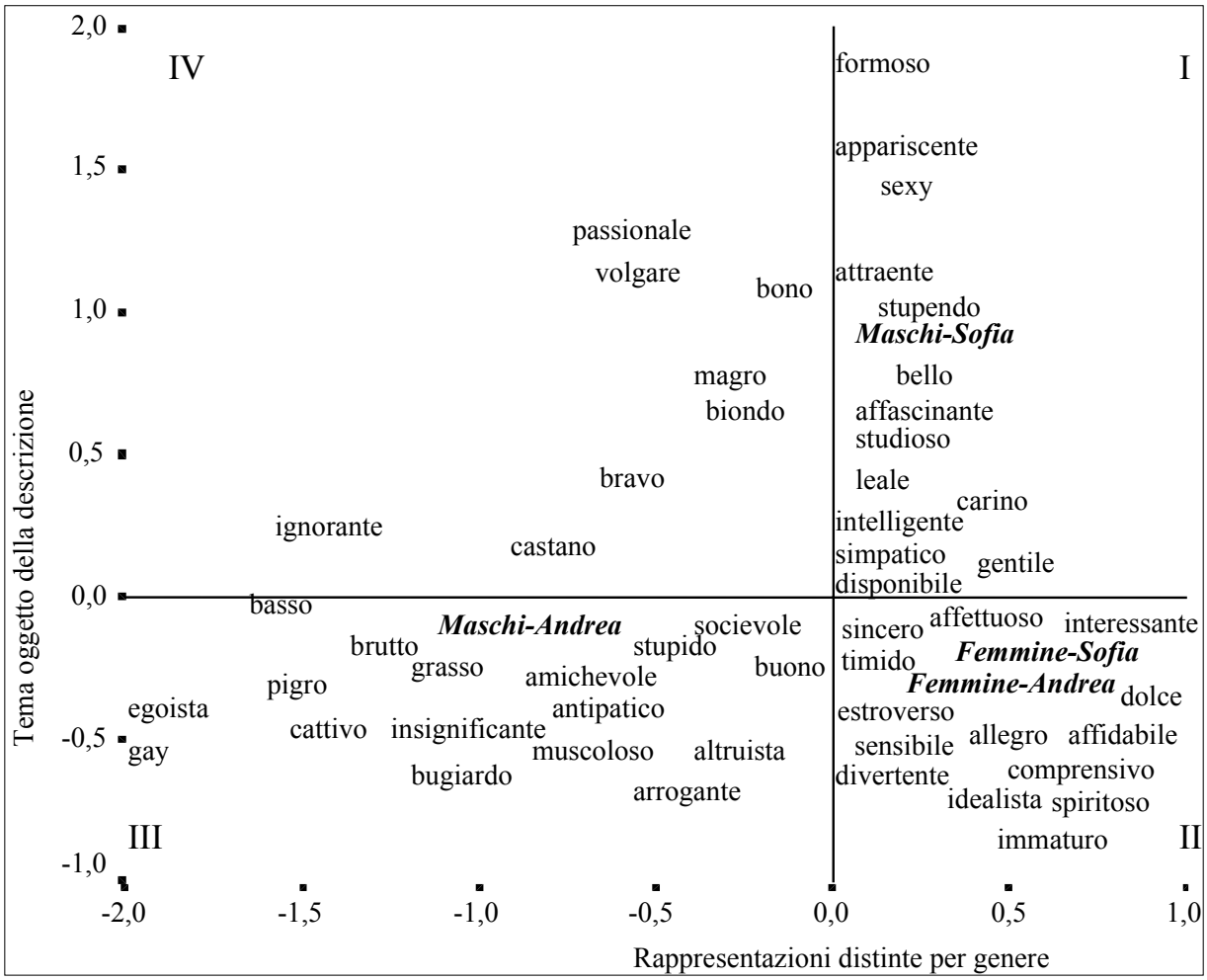


Figura 1. Piano fattoriale 1-2

Sembra altresì emergere con evidenza che nell'età considerata i maschi hanno, nei confronti delle femmine, anche quando amiche e non partner, un orientamento che tiene in grande considerazione la dimensione "estetico-corporea".

Poiché emerge una differenza tra dimensione relazionale, apprezzata tra le amiche e non tra agli amici, ai fini di un ulteriore approfondimento sarebbe interessante verificare i risultati con quelli espressi da domande analoghe riferite alla descrizione di un proprio partner reale o desiderato.

Bibliografia

Arcidiacono C. (1994). *Psicoanalisi e genere. Da donna a donna alla ricerca del senso di sé.* Franco Angeli.

Ammaniti M. e Dazzi N. (a cura di) (1991). *Affetti. Natura e sviluppo delle relazioni interpersonali*. Biblioteca di cultura moderna Laterza

Attili G. (2000). Le basi biologiche del comportamento sociale. La psicologia sociale evoluzionistica. In *Introduzione alla psicologia sociale*. SEAM, Formello Roma.

Aa.V.v. (a cura di Caterina Arcidiacono) (1991). *Identità genere differenza*. Franco Angeli.

Bolasco S. (1999). *Analisi multidimensionale dei dati*. Carocci.

Bolasco S., Morrone A. e Baiocchi F. (1999). A paradigmatic path for statistical content analysis using an integrated package of textual data treatment. In Vichi M. e Opitz O. (Eds), *Classification and data analysis. Theory and application*. Springer-Verlag.

Breen D. (1998). *The gender conundrum*. Routledge.

Burr V. (2000). *Psicologia delle differenze di genere*. Il Mulino.

Cicognani E. (2002). *Psicologia sociale e ricerca qualitativa.* Carocci.

Cipriani R. e Bolasco S. (a cura di) (1995). *Ricerca qualitativa e computer, analisi e applicazioni.* Franco Angeli.

D'Amico R. e Bimbi F. (a cura di) (1998). *Sguardi differenti*. Franco Angeli.

Eagly A. (1987). *Sex differences in social behaviour: a social role interpretation*. Hillsdale (NJ).

Farr R.M. e Moscovici S. (a cura di) (1984). *Rappresentazioni sociali.* Il Mulino.

Halpern D.F. (1986). *Sex Differences in Cognitive Abilities*. Lawrence Erlbaum Associates. Hillsade 1992.

Huston A.C. (1985). The development of sex typing: Themes from recent research. *Development Review*, vol. (5) : 1-17.

Lloyd B. (1994). Differenze di genere. In (a cura di Moscovici) *La relazione con l'altro.* Cortina.

Longobardi M. (2002). *Rappresentazioni di genere in adolescenza.* Tesi di Laurea in Psicologia di comunità, Seconda Università degli Studi di Napoli, Facoltà di Psicologia, relatore Prof.ssa Arcidiacono C.

Moscovici S. (1984). *Social Rappresentations.* Cambridge University Press.

Moscovici S. (a cura di) (1997). *La relazione con l'altro*. Raffaello Cortina Editore.

Mitaritonna A. e Nicolli S. (1998). *Femminile e maschile - percorsi alla scoperta della differenza*. AIED.

Nunziante Cesàro A. (1998). Identità di genere e adolescenza. *Psicologia clinica dello sviluppo*, vol. (1). Il Mulino.

Palmonari A. (1989). *Processi simbolici e dinamiche sociali*. Il Mulino.

Stoller R.J. (1968). *Sex and Gender*. Aronson.

L'Analisi Emozionale del Testo (AET) : un caso di verifica nella formazione professionale

Renzo Carli, Francesca Dolcetti, Nadia Battisti

Laboratorio di Analisi del Testo – Corso di Laurea in Scienze e Tecniche Psicologiche
per l'Intervento Clinico per la Persona, il Gruppo e le Istituzioni –
Facoltà di Psicologia 1 dell'Università "La Sapienza" di Roma –
Via dei Marsi 78 – 00185 Roma – Italia
renzo.carli@uniroma1.it, francesca.dolcetti@uniroma1.it, nadia.battisti@uniroma1.it

Abstract

This work proposes to make explicit those connections between tools and models which point out, in a given text, cultural models within a social context. Text Emotional Analysis (TEA) analyses texts within researches and interventions. Therefore, text as behavior, as proposal of relationship, as mode of treating emotional collusive dimensions into different contexts; language as organizer of relationships rather than detector of emotions and significance. Tracks of those representations at the base of decisional processes, which find orientation within the complexity of the individual-context relationship, are the *dense words*; those having strong emotional polisemy, here elaborated through their co-occurrances. We present a work on the verification of formation as example of new instruments of monitoring and of validation of procedures. The object of this verification, are the cultural models within the professional formation of students of a University psychology course, related to the demand of psychological services surveyed in their future field of practice in the Lazio region.

Riassunto

In questo lavoro esplicitiamo nessi fra strumenti e modelli per rilevare in un corpus testuale modelli culturali in un contesto. L'Analisi Emozionale del Testo (AET) analizza testi entro ricerche e interventi, testo come comportamento, modo di organizzare le dimensioni emozionali collusive entro i contesti, linguaggio come organizzatore di relazioni piuttosto che rivelatore di emozioni e di senso; le tracce delle rappresentazioni che fondano i processi decisionali, ed orientano nella complessità della relazione individuo-contesto, sono le *parole dense*, quelle a forte polisemia emozionale, qui elaborate attraverso la loro co-occorrenza. Presentiamo un lavoro sulla verifica come esempio sia dell'uso dell'analisi testuale secondo nuovi modelli di lettura dei processi sociali che della creazione di nuovi strumenti per il monitoraggio e la validazione di prassi. Oggetto di questa verifica sono i modelli culturali della formazione professionale degli studenti di una facoltà di psicologia, sul versante dell'offerta, messi in relazione con quelli della domanda di servizi psicologici, rilevati presso la loro futura utenza nella regione Lazio.

Parole chiave: analisi emozionale del testo (AET), ricerca-intervento, verifica, formazione professionale, immagine, domanda e offerta, cultura locale, modelli culturali, inconscio, parole dense, etimologia, analisi delle co-occorrenze, cluster analysis, Alceste.

1. Introduzione

Presentiamo un caso di uso dell'Analisi Emozionale del Testo (AET) in un intervento di verifica della formazione del corso di Psicologia Clinica (2000-2001), della Facoltà di Psicologia dell'Università "La Sapienza" di Roma. Questo intervento si è avvalso degli orientamenti emersi da un precedente lavoro di sviluppo dell'immagine dello psicologo nella popolazione della regione in cui ha sede la Facoltà; che la stessa Cattedra di Psicologia Clinica ha realizzato su committenza dell'Ordine degli Psicologi del Lazio (1999).

Entrambi sono lavori di ricerca-intervento sui contesti entro cui si muove la psicologia, e trattano gli articolati modelli culturali che organizzano, da un lato, le rappresentazioni che orientano gli studenti verso la loro futura professione, dall'altro, le domande che gli interlocutori sociali fanno allo psicologo, orientate dall'immagine che hanno di questa professione. L'ipotesi è che le risorse di una professione, dunque anche quella psicologica, si fondino sull' acquisire competenze a leggere il contesto entro cui prendono corpo domanda e offerta; se questa relazione si organizza a partire dalle rispettive rappresentazioni, sembra utile conoscere e sviluppare quelle caratteristiche in grado di aumentare la potenzialità di questo incontro. Da questa esplorazione l'Ordine degli Psicologi ha attivato strategie di sviluppo per l'immagine, la Facoltà di Psicologia ha avviato misure riorganizzanti la formazione.

Introduciamo di seguito alcuni elementi teorico-metodologici, così come i vari passi operativi, del metodo dell'AET: la raccolta del materiale testuale, la preparazione del corpus da mettere in analisi, la preparazione del vocabolario utilizzato per l'analisi (riduzione e scelta, inclusione/esclusione, delle forme grafiche presenti nel dizionario prodotto dal programma Alceste), l'interpretazione dei risultati dell'analisi, il loro utilizzo nell'intervento; tutte operazioni concepite entro uno stretto riferimento alle questioni e agli obiettivi posti dalla committenza.

2. Elementi teorico-metodologici dell'Analisi Emozionale del Testo

L'Analisi Emozionale del Testo (AET) è un metodo di ricerca-intervento sviluppato nell'ambito della psicologia dall'esigenza di avere strumenti funzionali nel lavoro psicologico clinico e psicosociale, nell'ottica di conoscere e supportare, in una ampia serie di campi di applicazione, lo sviluppo delle relazioni individuo-contesto.

2.1. *Perché ci interessa la dimensione emozionale veicolata dal testo?*

Riteniamo che la nostra mente funzioni entro due modalità: il *modo di essere inconscio* ed il suo *modo cosciente*. Il modo di essere inconscio ha come caratteristiche la condensazione, lo spostamento, l'assenza di negazione, l'assenza del tempo, la sostituzione della realtà esterna con quella interna; queste caratteristiche erano già individuate da Freud nella sua prima formulazione teorica. Mentre, il modo cosciente, sulla base della percezione e del pensiero, organizza e costruisce la realtà in relazioni, spaziali, temporali, definite e descrivibili, categorie di riferimento entro le quali è possibile per noi intervenire in maniera appropriata.

Il rapporto tra i due modi non è di esclusione quanto di reciproca sinergia: quello inconscio conservando le sue straordinarie qualità, di muoverci verso e, per così dire, di venire a patti con la realtà oggettuale, quello cosciente producendo differenze e stabilire relazioni tra oggetti. Qui si situa la dimensione emozionale, nei diversi modi della sua espressione, entro l'area che va dal modo di essere inconscio della mente al suo modo cosciente. Dove per inconscio non si intende un archivio di significati in un'ottica individuale, né una dimensione sovraindividuale come negli archetipi, quanto di un aspetto della doppia referenza di un processo mentale sempre presente. E' l'emozione che ci permette di istituire motivatamente le relazioni con gli oggetti del contesto, sulla base di simbolizzazioni affettive, quindi anche quelli sociali. Fra le prime sulle quali strutturiamo la nostra conoscenza troviamo: amico/ nemico, alto/basso, dentro/fuori, davanti/dietro.

Il costrutto di *collusione* è stato formulato per indicare la simbolizzazione affettiva del contesto da parte di chi vi partecipa; si tratta del processo di socializzazione delle emozioni, di un condividere emozionalmente le stesse simbolizzazioni affettive, come anche simbolizzazioni diverse ma complementari, entro un contesto partecipato e vissuto in comune.

Si vuole proporre una processo di sinergia fra i due costrutti di collusione e di *rappresentazioni sociali*, poiché ad entrambi è stata attribuita la funzione di ridurre e fronteggiare le dimensioni di estraneità poste dal contesto e quindi di contribuire a riorganizzare, nei processi di cambiamento, le categorie su cui fondare un adattamento. L'ipotesi è che la collusione possa interpretare genesi e motivazione delle rappresentazioni sociali, e che questa seconda possa essere assunta come riferimento metodologico per la rilevazione della collusione entro le relazioni del contesto. L'interesse è per la rappresentazione di una popolazione o un gruppo, intesa come scambio sociale che costruisce e condivide un oggetto comune, dove l'attenzione è al processo e agli elementi che generano una costruzione di relazioni, distinguendola dalla rappresentazione come fenomeno individuale, atteggiamento o opinione che sia, e dove il significato viene ricercato nel suo contenuto.

Questa prassi di lavoro utilizza un forte parallelismo fra l'approccio indiziario, quel modo di cercare tracce proprio della semeiotica e della clinica, e le opportunità offerte dalla statistica, con l'analisi delle corrispondenze multiple e la cluster analysis, e dall'informatica, con la predisposizione di software che permettono di trattare testi e grandi quantità di dati.

In quest'ottica l'uso di Alceste è motivato dalle prossimità che riteniamo esserci fra le ipotesi su cui si fonda l'AET e quelle del programma, nell'ottica che ci interessi evidenziare le relazioni che attengono alla produzione del testo e quindi anche quelle che riflettono le modalità di esprimersi della dimensione emozionale.

2.2. *Come pensiamo di rilevare la dimensione emozionale nel testo?*

Ricordiamo che lo psicologo clinico, sia nella consulenza per lo sviluppo individuale, sia in quella organizzativa, lavora attraverso le parole.

Attraverso il parlare, attraverso la narrazione, noi abbiamo due tipologie di effetti nel medesimo tempo: sotto il profilo pragmatico produciamo una costruzione di senso, quindi di coerenza tra parti e con ciò realizziamo un atto comunicativo, che ognuno di noi riconosce come intenzionale; dall'altro lato trasformiamo in "contesti di parole" la simbolizzazione affettiva e costruiamo emozionalmente una relazione con il contesto locale a cui la narrazione è diretta. Parola e testo sono da trattare in questo in questo caso come *atti* linguistici, come comportamenti, che producono un effetto nella relazione, in quanto proposte emozionate che implicano attese da parte di chi li produce e risposte da parte di chi interloquisce.

Con l'analisi emozionale si vuole cogliere la proposta di relazione, che la narrazione mira a produrre, la sua valenza comunicativa; il testo allora è sempre trattato in quanto iscritto dentro una relazione, e per questo non trattabile se non in stretto riferimento al contesto di un intervento che può essere già in atto o che si vuole sollecitare.

Così l'AET si differenzia sia da lavori di analisi del testo che si rivolgono alle dimensioni grammaticali, sintattiche, semantiche, sia da un approccio narratologico, che tende a cogliere la struttura e la coerenza tra le parti della narrazione, valutandone aspetti come la congruità, l'adeguatezza, la comprensibilità, la ricchezza di elementi, rispetto ad un qualche modello .

Ma quali parole costituiscono quei "contesti di parole" che veicolano la dimensione emozionale del testo?

Con Alceste Reinert propone l'idea di basare l'analisi, e così definire i differenti contesti lessicali del corpus, su di un vocabolario costituito da parole chiamate *piene* (nomi, verbi, aggettivi, alcuni avverbi). L'Autore, eliminando dall'analisi le parole strumentali necessarie

alla sintassi, già fonda l'analisi del testo fuori dall'obiettivo di analizzare la dimensione convenzionalmente intenzionale dell'atto linguistico.

Fin da quando l'AET ha iniziato ad usare programmi di analisi testuale, è emersa l'esigenza di raccordare l'uso del software con la possibilità di far emergere le dimensioni collusive della relazione individuo-contesto, entro una teoria della tecnica che tenesse in connessione gli strumenti con l'obiettivo del loro uso. Operativamente ha significato, ad esempio, intervenire sul vocabolario attraverso la messa in analisi di quelle parole che nel nostro lavoro individuiamo essere degli indicatori delle dimensioni collusive, questo perché esse rivestono un ruolo prioritario entro le classi di parole identificate dall'analisi.

Queste parole sono state chiamate *parole dense*, parole che più di altre veicolano le componenti emozionali del testo; parole che a differenza di altre hanno meno bisogno del contesto linguistico per poter istituire una relazione con il contesto locale. Alcuni esempi: le parole *bomba*, *madre*, *superare*, *viaggiare,* sono in grado di veicolare emozioni anche senza che siano dentro un contesto linguistico. La parola *andare* di per sé, invece, non ha una particolare densità emozionale, ha bisogno di un contesto linguistico, cioè di un'altra parola per significarci qualcosa di emozionale; questa parola può essere *via*, allora *andare_via* diventa una parola che suscita una emozione.

2.3. *Cultura Locale*

Chiamiamo *Cultura Locale* l'insieme quei processi collusivi, sui quali si pensa di intervenire con il lavoro psicosociale, (siano essi rilevati con l'AET e/o con altri metodi più complessi, come ISO-Indicatori di Sviluppo Organizzativo, qui solo accennati). Questa si articola al proprio interno in "sottoculture", che chiamiamo *Repertori Culturali* (RC) e che corrispondono alle classi di parole o ai cluster individuati attraverso l'analisi statistica.

3. La verifica della formazione professionale

Presentiamo il lavoro di verifica della formazione professionale degli studenti che hanno partecipato al corso di Psicologia Clinica nell'anno accademico 2000-2001, si tratta di una delle discipline che viene seguita al 4° e penultimo anno di università. Questo lavoro di verifica della formazione è proseguito negli anni accademici successivi ed è stato adottato anche da altri corsi della Facoltà. La sua attivazione sta incrementando una conoscenza e una trasformazione dei modelli della formazione professionale, anche nell'ottica di sostenere la riforma universitaria introdotta di recente in Italia. Daremo conto delle ipotesi che abbiamo fatto sull'evolversi dei modelli interni al gruppo di studenti, per indicare quelli che riteniamo maggiormente in grado di accogliere e sostenere gli aspetti di sviluppo del contesto e di trattarne gli aspetti critici, in una logica di verifica sia interna al processo formativo, sia esterna, rispetto alle rappresentazioni sulla figura dello psicologo attive nel conteso locale. Questo secondo aspetto della verifica sarà affrontato in modo specifico nel capitolo 4.

3.1. *Obiettivo: la verifica come azione formativa*

L'obiettivo della verifica è stato di rilevare e orientare una trasformazione dei modelli culturali della professione condivisi dagli studenti che hanno partecipano al corso, poiché questi hanno la funzione di sostenere e orientare l'azione sia durante l'esperienza formativa, che nell'avvio alla professione.

Vorremmo differenziare questo indicatore di verifica della formazione da altri, come l'acquisizione delle teorie o delle tecniche, che sono tradizionalmente verificabili con test e prove pratiche.

Sono state effettuate due rilevazioni: all'inizio e alla fine del corso. L'AET della prima rilevazione, che ha consentito una conoscenza iniziale dei modelli culturali della professione con la quale gli studenti si avviavano al corso, è stata utilizzata per orientare l'intervento formativo attraverso un processo di riorganizzazione della relazione tra docente e studenti. Allo stesso modo, è stato utilizzato il lavoro finale di verifica qui presentato, che ha trattato i due scritti (inizio e fine) in un incontro di restituzione proposto agli studenti, con l'obiettivo di verificare assieme il percorso fatto e segnalarne le dimensioni critiche e di sviluppo.

3.2. *La produzione del testo: costruzione della committenza*

Gli studenti hanno partecipato alla rilevazione in due momenti del corso: alla settima lezione e alla terz'ultima. Un'attenzione particolare è stata dedicata ai criteri che ispiravano la raccolta dei testi: costruire la loro motivazione ad un'implicazione nel lavoro di verifica, ciò è stato fatto entro il setting delle prime sette lezioni; sostenere una produttività testuale per associazione di idee, proponendo agli studenti la seguente traccia tematica e invitandoli a scrivere quello che veniva loro alla mente: "Descriva cosa pensa di fare come psicologo clinico: quali attività pensa di svolgere, in quali ambiti, con quali risultati e quale pensa sia la domanda da parte di possibili clienti?". Alla prima rilevazione hanno partecipato il 43.5% degli studenti frequentanti (57 studenti su 131), alla seconda la partecipazione è stata del 35.6% (26 studenti su 73); poiché il concetto di cultura prescinde dal singolo individuo e riguarda l'intero gruppo che condivide il contesto analizzato, questa diminuzione, che è all'interno di un trend tipico della Facoltà, riteniamo non invalidi i risultati della verifica.

3.3. *Preparazione del corpus*

Il corpus è stato composto con tutti i testi raccolti; qui le unità iniziali sono state distinte con la variabile illustrativa del tempo della rilevazione (SCR_1 e SCR_2) e con la distinzione data dai singoli testi di ciascuno studente. Nel corpus abbiamo operato modifiche:

a) abbiamo formato alcune politematiche, ad esempio *analisi_della_domanda*, che ci interessava fosse trattata nell'analisi come un'unica forma, poiché sta ad indicare un metodo di intervento psicologico oggetto del corso e dunque emozionalmente rilevante ai fini del lavoro di verifica; di contro, in un altro contesto, queste tre parole messe assieme non avrebbero lo stesso significato e non verrebbero polirematizzate;

b) è stato disambiguato il senso diverso di alcune forme omografe; ad esempio, la parola *facoltà* ha due sensi: *facoltà mentale e fisica* e la *facoltà universitaria*. Nel nostro caso, era rilevante differenziare queste modalità d'uso della parola, che consideriamo entrambe emozionalmente dense in questo contesto di verifica. Più scontato è l'esempio della forma *legge*, terza coniugazione del presente indicativo del verbo leggere, ovviamente differenziata nel corpus dalla forma *legge*, laddove questa indicava la norma.

3.4. *Dizionario: lessematizzazione e scelta delle forme accettate/rigettate nell'analisi*

Una volta pronto il corpus, il passo successivo è quello di far produrre ad Alceste il dizionario del corpus, nell'opzione fornita dal programma costituita dalle sole forme grafiche e dalle relative occorrenze. Abbiamo poi realizzato manualmente la lessematizzazione e la scelta delle forme da accettare/rigettare nell'analisi. Per la lessematizzazione, il criterio utilizzato è ridurre tutte le forme del lessema ad una medesima forma. Ad esempio, il verbo *sbagliare* (con tutti i suoi tempi e le sue coniugazioni), l'aggettivo *sbagliato*, il sostantivo *sbaglio*, e quindi le loro forme al singolare, al plurale, al maschile e al femminile, sono ridotti alla medesima forma *sbagl*<. Rigettiamo, così come fa Alceste, tutte le parole strumentali ed i numeri. Procediamo poi ad un'ulteriore scelta quella delle *parole dense*, che caratterizzano in

modo particolare l'AET. Tra le parole che in questa analisi sono state rigettate, diversamente dalla proposta di Alceste, diamo di seguito degli esempi: *ambito*, *andare*, *espressione*, *fare*, *forma*, *intendere*, *mettere*, *usare*. Queste sono rintracciabili in ogni caso nel report finale dato dal programma, perché contrassegnate da un asterisco (*) entro le diverse classi.

3.5. *La Cultura Locale degli studenti del corso*

Nell'analisi le rappresentazioni professionali, rilevate presso gli studenti del corso, sono emerse organizzate in 3 classi o Repertori Culturali (RC), come si può vedere nello spazio fattoriale di seguito presentato (Fig. 1).

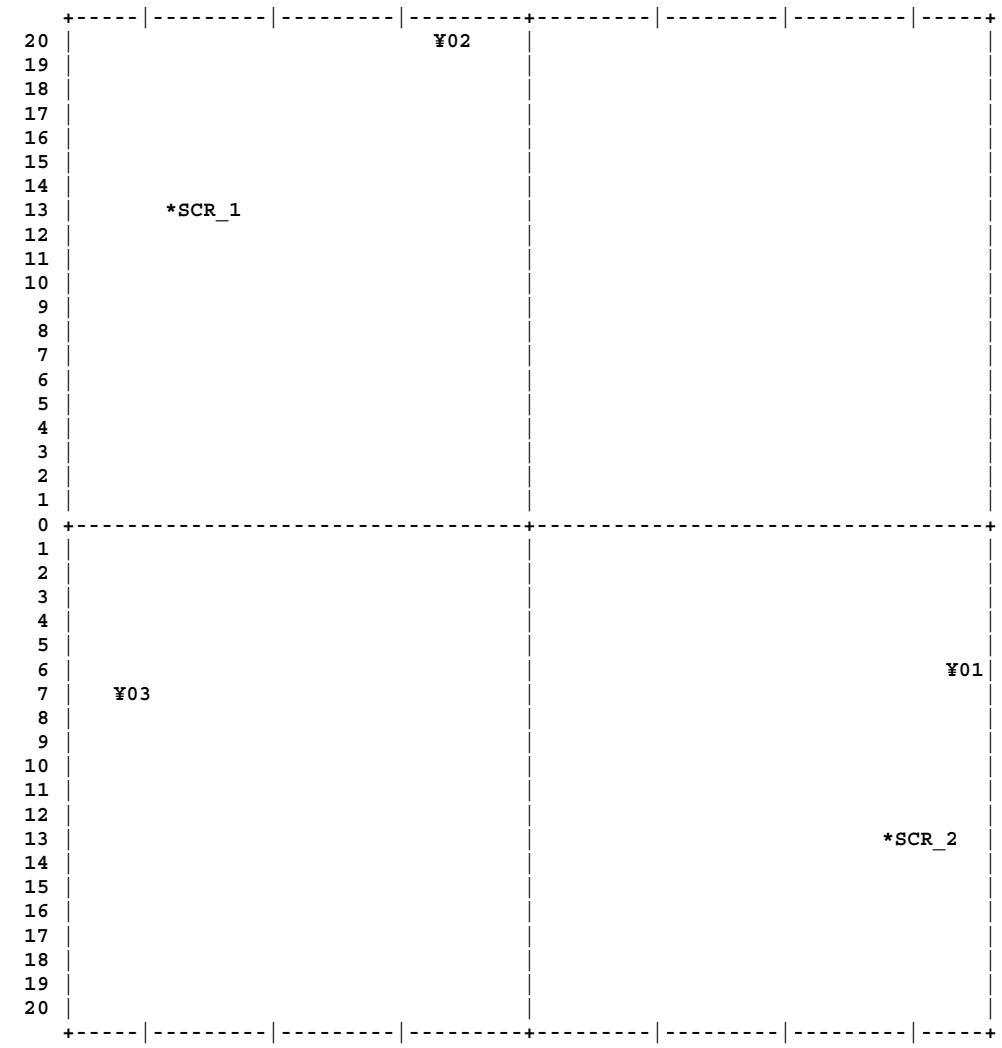


Figura 1.

3.5.1. *Interpretazione dei Repertori Culturali: restrizione della polisemia emozionale*

Per l'interpretazione di ciascun RC, utilizziamo come criterio la restrizione della polisemia emozionale nell'incontro progressivo tra ciascuna parola densa entro la classe, partendo dalla parola con χ^2 più alto, quindi più centrale nel cluster, e procedendo di seguito nell'incontro con le parole con χ^2 più basso. Un modo produttivo per recuperare il massimo della polisemia emozionale delle parole è l'utilizzo della loro etimologia.

Nella fig. 2 le forme ridotte delle 3 classi, la variabile categoriale con due modalità, il loro χ^2.

χ^2	**RepertorioCulturale 2** Nombre d'u.c.e.: 122 soit: 21.75%	χ^2	**RepertorioCulturale 3** Nombre d'u.c.e.: 178 soit: 31.73%	χ^2	**RepertorioCulturale 1** Nombre d'u.c.e.: 261 soit: 46.52%
31,34	sbaglia<	37,72	facoltà	62,64	contest<
31,34	sper<	36,25	ide<	57,34	interven<
25,51	sofferen<	33,16	laure<	43,10	svilupp<
24,23	comportament<	32,51	dopo	43,00	organizza<
24,23	tribunal<	27,16	teori<	41,46	costru<
20,83	viver<	27,00	esam<	31,54	grupp<
20,62	giust<	26,38	cura<	30,08	prod<
20,30	disagi<	26,38	iscri<	22,04	aziend<
17,63	soddisfa<	25,47	psicologia_clinica	21,38	process<
15,27	famigli<	25,25	lezion<	18,13	convive<
11,75	riusc<	25,01	psicoanali<	16,94	dinamic<
11,02	chied<	21,47	studen<	16,50	collusi<
10,85	infan<	20,34	prof_del_corso	16,50	relazional<
10,85	superficiali<	17,77	scel<	16,30	competen<
10,28	bambin<	17,18	semplic<	14,76	cambiament<
10,06	giuridico	16,58	impara<	12,39	falliment<
10,06	minor<	15,07	concret<	12,26	pensier<
61,07	**SCR_1**	**18,23**	**SCR_1**	**109,13**	**SCR_2**

Figura 2.

3.5.2. *Repertorio Culturale 2*

In questa classe la parola con χ^2 più alto è *sbagliare*, che deriva originariamente da abbagliare: "uscire dai raggi di luce", l'idea è di uscire al di fuori di un percorso prestabilito da qualcosa (le leggi della fisica). Ipotizziamo qui un evento critico che viene vissuto con un senso di inadeguatezza. Vediamo come questo viene affrontato nell'incontro con la successiva parola *sperare*: tendere verso, aspirare a, riponendo fiducia in qualcosa o qualcuno; si mantiene qui una posizione down dove il recupero di una qualche padronanza è affidata ad un esterno idealizzato e potente. Il successivo incontro è con la parola s*offerenza* che origina dal sopportare: ancora una condizione problematica di sottomissione. Si sta delineando una relazione con il contesto fondato sulla ricorrenza idealizzazione-impotenza, di confronto fra un modello e lo scarto da questo. L'incontro successivo è con le parole *tribunale* e *comportamento*. Il *tribunale* trae origine da *tribù*, luogo che organizza l'essere dentro o fuori un gruppo, suo rappresentante era il tribuno, magistrato o funzionario che esercitava la giustizia parlando da un alto scranno. *Comportamento* indica qualcosa di esternabile e di visibile, che in quanto risultato può essere valutato in rapporto a dei modelli; inoltre rappresenta un costrutto tipicamente oggetto della psicologia, dunque con un significativo investimento emozionale. *Tribunale* e *comportamento* delineano una formazione orientata da adesione a norme, ciò che c'è di consolidato e già tracciato, dove l'attenzione è ad elementi ostensibili, stati finali piuttosto che processuali degli eventi. Entrambi si rivolgono agli individui e rappresentano ancoraggi dove si recupera una padronanza: la certezza procurata dal tribunale e la verificabilità del comportamento. Un assetto che ben risponde ad un mandato sociale nei confronti di chi sbaglia, che propone un ***modello di relazione fondato sul controllo***. Più avanti troviamo il *disagio* "non giacere presso" e *famiglia*, *infanzia*, *bambini*, *minori*, come utenti "deboli", coloro che "vengono mandati" e che non esprimono direttamente allo psicologo delle domande di servizio.

3.5.3. *Repertorio Culturale 3*

Il primo incontro è tra *Facoltà* ed *idee*. *Facoltà* come capacità, potenzialità, potere, ma anche l'insieme di discipline che organizzano l'università, dal latino *făcilis*: facile, agevole, adatto, propenso. *Idee* dal greco *idêin* "vedere". Gli studenti attraverso la *Facoltà* "stanno a vedere", osservano, più che di mettersi in gioco. Il successivo incontro è con *laurea*, dal latino *lāureu*: di alloro, allusione alla corona d'alloro ornamento di dei e uomini vittoriosi, c'è la tensione a raggiungere un obiettivo che al momento non è altro che l'essere primi. L'associazione con *dopo* organizza il rapporto con il contesto entro un tempo altro da quello attuale, come se il senso di onnipotenza, dove tutto è ancora possibile, abbia bisogno di un limite che lo renda sostenibile, mettendosi temporalmente fuori l'azione formativa già in corso. L'incontro con *teorie*, dal greco *theōría*, parola composta da *theá* cioè spettacolo e *horân*, osservare, conferma una posizione da spettatore piuttosto che da partecipante della formazione. *Esami*, dal latino *âgere* ed *ex-*, cioè pesare bene, è la dimensione di un radicamento che si comincia a sperimentare. Ma quali ne sono i punti di riferimento? L'incontro con *cura*, è sia la posizione di essere ancora presi in carico che uno degli aspetti dell'agire professionale in appoggio alla professione medica, che tradizionalmente ha un forte mandato sociale, entro un ***modello di simbolizzazione della relazione alto/basso***. Il desiderio di sentirsi dentro, *in-scritti* nella *psicologia clinica*, rafforzato dal riferimento alla *psicoanalisi*: stereotipo professionale dello psicologo clinico, che vagheggia qui la partecipazione ad un gruppo di appartenenza fortemente strutturato, capace di fornire un ancoraggio emozionale mitico alla professione, una formazione molto selettiva e lunga, possibile solo dopo l'università. Torna il ***modello dentro/fuori*** di una forte domanda di appartenenza, dove i soli clienti sono gli *studenti* stessi, in una formazione che sembra non avere mai fine.

3.5.4. *Repertorio Culturale 1*

La parola *contesti*, dal latino *contĕxtu,* tessere insieme, intrecciare, unire, dimensione di un fare che costruisce un legame, in questo repertorio incontra la parola *intervento*, dal latino *venīre ĭnter*, arrivare nel mezzo, la partecipazione attiva a fatti o a situazioni, come anche prendere la parola. Si associano due dimensioni di azione dentro uno spazio, dove la parola *sviluppo* specifica lo scopo e l'effetto dell'intervento: da *viluppo s-* che significa sciogliere un viluppo, un intreccio, ma anche far progredire, aumentare, incrementare, un'espansione, un potenziamento, nel rapporto con la formazione. La dinamica collusiva di questa classe è ricca di verbi. Abbiamo così *organizzare*, parola composta da *órganon* e *-izzare*. *Órganon* ha la stessa radice di *érgon*, lavoro, opera; il significato esteso sta ad indicare la parte che in un complesso adempie ad una precisa funzione coordinata con quella delle altre parti, significa anche coordinare, disporre i vari elementi in modo da raggiungere un fine. *-Izzare* significa agire in un certo modo. L'incontro con il verbo *costruire*, *strŭere con-*, edificare insieme, congiungere, formare un insieme organico dato da diversi elementi, implica un ***modello di rapporto con l'altro, l'estraneo***, come se la formazione avesse svolto un'attivazione emozionale diffusa rispetto alla dipendenza che caratterizza la posizione dello studente, un movimento di presa in carico di uno scopo che si organizza intorno alle parole successive: *gruppo*, *prodotto*. La parola *gruppo*, in particolare, ci fa pensare all'intervento psicologico, sia riferito ai clienti che agli studenti che lo scoprono come prodotto della formazione stessa in cui si è sviluppata una dimensione di appartenenza, già nel conteso universitario, risorsa importante per l'apprendimento. Pensiamo ad un ***modello di relazione fondato su un'azione di costruzione della relazione individuo-contesto***, in cui l'iniziativa e l'esplorazione costruite entro il percorso formativo hanno un'influenza sulla rappresentazione della professione.

3.5.5. *Risultati della verifica dell'azione formativa*

Secondo gli elementi che qui abbiamo potuto accennare, la posizione culturale degli studenti espressa nel primo scritto, prevalentemente rappresentata dal RC 2 (SCR_1 con χ^2 61,07), prefigura una professione psicologica ancorata a dimensioni normative, sulla base di un mandato sociale forte della preoccupazione per la sofferenza ed il deficit di individui deboli, probabilmente vittime dello scarso rispetto delle regole della convivenza: siamo nel modello dell'aiuto che lo psicologo condivide con altre professioni che si occupano della marginalità. Il RC 3 (SCR_1 con χ^2 18,23) ancora la sua visibilità sociale e la sua capacità di intervento al possesso di tecniche con una tradizione prestigiosa, buone in sé indipendentemente da una lettura del contesto dove nasce l'esigenza di cura, che ha ancora come specificità l'intervento sull'individuo. Nel RC 1 (SCR_2 con χ^2 109,13) emerge una posizione di esplorazione su nuovi oggetti sociali (azienda, produzione, convivenza) e su dimensioni tecniche non ancorate al bisogno di un potere/mandato predefinito, per trattare gli eventi critici del contesto, tessendo e organizzando relazioni fra le parti in gioco attraverso risorse del pensiero.

4. Lo sviluppo di strategie per l'immagine dello psicologo nella regione Lazio

Questo lavoro ha avuto l'obiettivo di individuare delle strategie di sviluppo della professione per l'Ordine degli Psicologi del Lazio, strategie basate sulla conoscenza delle componenti culturali dell'immagine, ossia della percezione di rapporto con l'altro da parte della popolazione che potenzialmente può rivolgersi alle prestazioni dello psicologo. Qui presenteremo solo la parte relativa alla conoscenza della Cultura Locale, realizzata in tre fasi, tralasciando le strategie di sviluppo individuate.

4.1. *Descrizione delle fasi e degli strumenti dell'indagine*

La prima fase, analogamente alla precedente sulla verifica, ha comportato un'esplorazione con AET, entro un campione per quote di 64 soggetti, rilevando le dimensioni emozionali con cui gli abitanti del Lazio organizzano le proprie conoscenze sullo psicologo.

Nella seconda fase è stato applicato uno strumento chiamato *quasi-questionario* (per approfondimenti vedi http://www.spsonline.it). La sua caratteristica è permettere l'individuazione dei linguaggi allo stato nascente accanto a quelli consolidati. Una "vecchia cultura", magari problematica e in cambiamento, si avvale di codici consolidati, ripetuti, di immagini e metafore del linguaggio comune mentre una "nuova cultura" che si sta affacciando risulta spesso senza parole, dandoci l'impressione di non esistere. Di qui l'utilizzo, nel quasi-questionario, di quelle parole dense, che nella fase precedente risultano avere un buon peso specifico all'interno dell'analisi, caratteristiche di quella popolazione e di quel contesto-problema; nonché dell'inserimento di elementi in grado di richiamare nell'intervistato l'espressione di specifici modelli culturali di rapporto con il contesto sociale Ne citiamo alcuni: le regole della convivenza, la posizione di cliente/utente nei confronti della Pubblica Amministrazione, l'integrazione nella propria zona e le previsioni di sviluppo e di successo della zona stessa, il sentimento di poter o meno influenzare il contesto. Il campione per quote è stato, in questa fase, di 393 soggetti. I dati sono stati trattati con l'analisi delle corrispondenze multiple e la cluster analysis, con SPAD.

L'obiettivo di costruire ipotesi piuttosto che verificarne, nell'interesse a recuperare tutte le rappresentanze, ha fatto scegliere un campione per quote non probabilistico, in una sorta di operazione di salvaguardia delle minoranze, nel nostro caso di natura culturale. Per variabili utilizzate (sesso, fasce d'età, ed area geografica) vedi il testo in bibliografia.

Nella terza fase il quasi-questionario, attraverso un'analisi discriminante, è stato ricondotto ad un *questionario snello* di rapida applicazione, con gli items più potenti della precedente fase, proposto ad un campione stratificato di 1064 soggetti, estratti casualmente e bilanciati in relazione alla proporzione della popolazione della regione. In questa fase sono stati acquisiti i dati quantitativi sulla Cultura Locale, segmentando il campione entro i Repertori Culturali emersi nella fase precedente; diversamente, se lo studio della variabilità dei soggetti fosse stata trattata segmentando preliminarmente il campione secondo variabili socio-demografiche, queste ultime avrebbero assunto la funzione generativa delle differenze rilevate.

4.2. *La Cultura Locale sull'immagine dello psicologo*

Elaboriamo ora gli elementi emersi nella seconda fase di questo lavoro. Una prima questione fondante: i differenti RC sono organizzati attorno modelli di rapporto con il contesto sociale, ciò significa che ognuna delle differenti componenti dell'immagine degli psicologi sono risultate strettamente correlate con una specifica rappresentazione culturale del territorio espressa dai cinque RC (per le illustrazioni vedi il testo in bibliografia).

4.2.1. *Il primo asse fattoriale: RC 4 la trasgressione delle regole e RC 3 il modello dell'aiuto alla sofferenza*

Sui due poli del primo asse fattoriale, prima *struttura* che genera la Cultura Locale, si situano i RC 4 e 3. Facciamo l'ipotesi che sia proprio il RC 4 a stimolare complessivamente l'organizzazione delle rappresentazioni di questo contesto e dunque anche del mandato sociale rivolto allo psicologo. Nel RC 4 infatti si è conferma la presenza, indicata da altre ricerche svolte sul territorio, di una minoranza che ha un'alta visibilità sociale: una parte della popolazione che cerca e persegue il proprio successo entro un contesto preteso senza vincoli né controlli; si tratta di soggetti incuranti e trasgressivi delle regole del gioco, pensate valide solo per gli altri che svalorizzano la qualità della vita del luogo in cui vivono e di cui non prevedono uno sviluppo, così come dell'ambiente, dei rapporti sociali, del senso civico dei cittadini, dello sviluppo economico e della politica. Questi contesti sono vissuti senza obiettivi produttivi, dove i modelli assunti sono quelli della competitività e del successo ad ogni costo, negando risorse nella relazione con l'altro, in una logica opportunistica e predatoria. Gli appartenenti a questa cultura assimilano a sé lo psicologo e ne diffidano perché un professionista pensato totalmente centrato, come loro, sui propri problemi di successo e di acquisizione di prestigio sociale, lo vedono come una figura che influenza e plagia i propri pazienti, lo associano al mago dalle competenze illusorie, andare dallo psicologo è alla moda, superfluo e capriccioso, secondo un cliché pretestuoso ed esibizionista. L'assimilazione nella trasgressione: un professionista che approfitta dei pazienti e dell'ansia che li ha portati da lui.

Nel RC 3, al polo opposto sul primo asse fattoriale, incontriamo una cultura fondata sulla fiducia nelle iniziative della Pubblica Amministrazione, sull'importanza della qualità dell' ambiente e delle regole della convivenza. Il sentimento dominante è la preoccupazione per le regole del gioco e per l'educazione al loro rispetto: è la risposta contrappositiva alla cultura precedente. Qui ci si rifugia nella famiglia e nei sistemi di controllo (scuola, sanità, forze dell'ordine appunto, e servizi di psicologia) modelli organizzativi nei quali si ha una fiducia cieca, vissuti come depositari delle regole della convivenza e regolatori dei comportamenti. L'immagine dello psicologo ha rilevanza come un *educatore* che opera nei contesti della scuola e della famiglia; la sua funzione essenziale è quella di *aiutare*, chi è in difficoltà, chi propone con il suo comportamento l'ipotesi di una sofferenza psichica, legata al deficit e alla devianza. La centralità della sofferenza bonifica simbolicamente il pericolo insito nei trasgressori e affida allo psicologo-educatore il potere di ricondurre sulla retta via sia chi agisce,

sia chi è vittima della violenza. E' la posizione di delega di un utente che riconosce allo psicologo, come alle istituzioni pubbliche, solo una regressiva rassicurazione; infatti costoro non pensano di utilizzarne i servizi, in vista di qualche proprio obiettivo di sviluppo, e le prestazioni sono destinate piuttosto ad altri, i *sofferenti* appunto.

4.2.2. *Il secondo asse fattoriale: RC 1 il modello della psicoterapia e RC 5 il modello difensivo dell'emarginazione della malattia mentale*

Rispetto ad una cultura della competitività senza regole ed irrispettosa dell'altro, nel RC 1 si reagisce con la paura del confronto sociale, sfiducia nella Pubblica Amministrazione e nelle iniziative che concernono la comunità. L'unico rifugio sicuro appare la famiglia dove, come alternativa all'assenza delle regole del gioco, c'è il controllo sostitutivo, dove la fiducia si trasforma in un familismo entro il quale ci si affida e ci si arrocca. La società è vista come una famiglia allargata, fondata su amicizie, conoscenza dei potenti, favoritismi. Questa cultura valorizza fortemente lo psicologo, che ha già consultato o al quale pensa di potersi rivolgere; esso è di sostegno alla famiglia, si occupa del disagio esistenziale e della tutela dei minori. Lo psicologo è identificato con lo psicoterapeuta, connotato come forte, il ricorso al quale sembra un ulteriore passaggio ad una dimensione protettiva, sostitutiva della famiglia, cui si indirizza una domanda di sostegno, non di incremento di competenze.

Nel RC 5 siamo al polo opposto del precedente; qui emerge fiducia nei confronti del contesto sociale, tradotta in un apprezzamento per l'affidabilità delle strutture di "controllo" (scuola, forze dell'ordine, ospedali, e servizi psicologici) luoghi entro i quali vigono delle norme, mentre viene svalutata l'Amministrazione locale e le iniziative che può intraprendere per la convivenza: Ambiente, rapporti sociali, offerta culturale, opportunità di lavoro e qualità dei servizi non sono ritenute aree rilevanti. Ci si sente impotenti nel far andar bene le cose e si sottovaluta tutto ciò che non è specificamente dedicato al controllo sociale. Siamo confrontati con una cultura dell'impotenza che chiede allo psicologo, assimilato allo psichiatra ed al sacerdote, solo controllo nei confronti della devianza vissuta come pericolosa. Lo psicologo assimilato all'area della sanità, si occupa di malati mentali entro una visione organicista che mette al centro il cervello e l'alterazione organica ma non la mente.

4.2.3. *RC 2 la fiducia nel futuro: siamo al polo positivo del terzo fattore*

Qui, per la prima volta, emerge chiaramente una *cultura non anomica*, differente da quelle sino ad ora analizzate, ove la fiducia nel pubblico e nello sviluppo del contesto si fonda su dimensioni di competenza; una minoranza (laureati, dirigenti, età 36-55 anni) in grado però di esprimere una forte pressione culturale; essi si pensano quali clienti della Pubblica Amministrazione, vista quale responsabile di un prodotto, non come gestore degli adempimenti dei propri utenti. C'è speranza nella qualità della vita, il luogo dove si abita viene vissuto come amichevole e facilitante la convivenza. In questo RC lo psicologo è un consulente per lo sviluppo: affronta i problemi della convivenza e che derivano dai conflitti entro le organizzazioni, interviene entro le problematiche sociali, comprende i processi organizzativi. Oggetto dell'intervento psicologico è il pensare e chi si rivolge allo psicologo, in questa cultura, lo fa come persona adulta, senza dipendenza acritica, attendendosi una consulenza per lo sviluppo della convivenza e per l'efficacia dei sistemi sociali, un'azione nell'ambito della relazione tra le persone ed il loro contesto, più che le problematiche dei singoli individui.

5. Conclusioni: alcune relazioni significative fra i due lavori presentati

Abbiamo individuato, attraverso l'immagine, la rappresentazione di un contesto con una forte ipotesi di trasgressione delle regole della convivenza che dunque orienta le attese nei

confronti dei servizi psicologici, verso un sostanziale ruolo di controllo sulla sofferenza, sulla devianza, sul disagio esistenziale, sulla malattia mentale. E' a questa domanda che gli studenti nel primo scritto si preparavano ad aderire, in una logica dove le proprie future prestazioni sono spesso assimilabili a quelle di altre professioni, quella dell'aiuto e quella medica. Nella popolazione del Lazio, è piuttosto una minoranza, potenzialmente interessante perché configurabile come opinion leader, che manifesta interesse per competenze specialistiche centrate sullo sviluppo della relazione fra gli individui e i loro contesti, e che caratterizzano maggiormente il secondo scritto degli studenti. Questa cultura orienta lo sviluppo della professione verso interventi che possono passare, ad esempio, dall'occuparsi del disagio scolastico dei singoli studenti, allo sviluppo della cultura del gruppo-classe come risorsa per l'apprendimento; dall'intervento sul burn-out degli operatori, alla riorganizzazione delle funzioni di accoglienza di un'agenzia sanitaria; dal ridurre l'aggressività individuale del personale, al ripensamento della relazione tra gli eventi critici, le trasformazioni del contesto e gli obiettivi produttivi di un'azienda.

Bibliografia

Bolasco S. (1999). *Analisi Multidimensionali dei dati.* Carocci.

Carli R. (1990). Il processo di collusione nelle rappresentazioni sociali. *Rivista in Psicologia Clinica*, vol. (4) : 282-296.

Carli R. e Paniccia R.M. (1999). *Psicologia della formazione.* Il Mulino.

Carli R. e Salvatore S. (2001). *L'immagine della psicologia*. Edizioni Kappa.

Carli R. (2001). *Culture Giovanili*. Il Mulino.

Carli R. e Paniccia R.M. (2002). *L'analisi emozionale del testo.* Franco Angeli.

Carli R. e Paniccia R.M. (2003). *L'analisi della domanda.* Il Mulino.

Cipriani R. e Bolasco S. (1995). *Ricerca Qualitativa e Computer*. Franco Angeli.

Ginzburg C. (1988). *Miti emblemi spie.* Einaudi.

Jodelet D. (1992). *Le rappresentazioni sociali.* Liguori.

Mazzara M. (2002). *Metodi qualitativi in psicologia sociale.* Carocci.

Moscovici S. (1989). *Psicologia sociale.* Borla.

Palmonari A. (1987). *Processi simbolici e dinamiche sociali.* Il Mulino.

Reinert M. (1993). *Les mondes lexicaux et leur logique à travers l'analyse statistique d'un corpus de récits de cauchemers.* In Cipriani R. e Bolasco S. (1995), *Ricerca Qualitativa e Computer*. Franco Angeli.

Linguaggio, ideologia e categorizzazione sociale : un'analisi psicologico sociale del documento di rivendicazione dell'attentato a Marco Biagi

Antonio Chirumbolo, Alessandra Areni

Dipartimento di Psicologia dei Processi di Sviluppo e Socializzazione
Università di Roma "La Sapienza" – Via dei Marsi 78 – 00185 Rome – Italia
chirumbolo@uniroma1.it, alessandra.areni@uniroma1.it

Abstract

The aim of this paper is to investigate, in a social psychological perspective, the rhetoric strategy behind the document of rivendication of the attempt on Marco Biagi's life, through the application of the Correspondence Analysis (CA). The material used is represented by the document of rivendication of the attack made by the RED BRIGATES on March the 19th 2002. Results pointed out four interpretable factors. The first two factors were loaded by few key words that stated instruments and goals of the organization. The third and fourth factors presented, instead, less obvious and more interesting contrapositions. The third factor can be understood as a syntagma that expresses the goal and the political proposal of the ingroup (i.e. REVOLUTION, DICTATORSHIP, PROLETARIAT), opposed to the political and economical goal of the outgroup (i.e. REFORM, LABOUR, MARKET). The fourth factor referred to the "political actors", contrasting words that define and contest the outgroup (GOVERNMENT and UNION) with words that define and contest the ingroup (i.e. RED BRIGATES). From a social psychological point of view, the rhetoric strategy emerged from the analysis of the document can be understood according to "Self-Categorization Theory" (Reicher, 1996 ; Reicher & Hopkins, 1996 ; Turner, 1982).

Riassunto

L'obiettivo di questo contributo è quello di studiare, in un'ottica psicologico sociale, la strategia retorico-discorsiva del documento di rivendicazione dell'attentato a Marco Biagi attraverso l'applicazione dell'Analisi delle Corrispondenze (AC). Il materiale è rappresentato dal documento di rivendicazione dell'attentato compiuto dalle BR il 19 Marzo 2002. I risultati hanno evidenziato quattro fattori latenti interpretabili. I primi due fattori sono risultati totalmente saturati da poche parole chiave che fissano mezzi e fini politici dell'organizzazione. Il terzo e quarto fattore hanno presentato, invece, delle contrapposizioni meno ovvie e particolarmente interessanti. Il terzo fattore può essere interpretato come un sintagma che esprime l'obiettivo e la proposta politica dell'ingroup (RIVOLUZIONE, DITTATURA, PROLETARIATO) contrapposto all'obiettivo economico-politico perseguito dall'outgroup (RIFORMA, MERCATO, LAVORO). Il quarto fattore sembra far riferimento agli "attori politici" implicati, contrapponendo parole che definiscono e contestualizzano gli "avversari", l'outgroup (ESECUTIVO e SINDACATO) a parole che definiscono e contestualizzano l'ingroup (BRIGATE ROSSE). Da un punto di vista psicologico sociale, la strategia retorica emersa dall'analisi del documento può essere interpretata alla luce della "Self-Categorization Theory" (Reicher, 1996 ; Reicher e Hopkins, 1996 ; Turner, 1982).

Parole chiave: linguaggio, ideologia, categorizzazione sociale, Brigate Rosse.

1. Introduzione

Le produzioni discorsive ed il linguaggio risultano essere tra gli ambiti d'elezione per lo studio delle ideologie (Augoustinos, 1998 ; Billig, 1991 ; Van Dijk ; 1998), delle rappresenta-

zioni della politica (Sensales, Chirumbolo e Areni, 2002 ; Sensales, Chirumbolo, Areni e Bettini, in stampa ; Sensales, Chirumbolo, Areni e Kosic, 2002), nonché delle categorie e delle identità sociali implicate nella comunicazione politica (Reicher, 1996 ; Reicher e Hopkins, 1996). In una prospettiva psicologico sociale di analisi del linguaggio politico, l'attenzione viene posta su colui che parla (o scrive) in quanto attore politico e sul modo in cui "motivazioni, scopi, emozioni, cognizioni, strategie orientano il soggetto nella produzione del discorso" (Catellani, 1997 : 133). Particolare attenzione, quindi, è rivolta alla struttura sintattica e al processo retorico-argomentativo intrinsecamente presente nella comunicazione politica (Billig, 1991). In questo senso, inoltre, è plausibile sostenere che anche una situazione di monologo (come per es. un discorso o un testo scritto) può essere considerata e studiata con i criteri del dialogo, in quanto in essa sono sempre presenti un intento retorico e degli assunti condivisi con un ipotetico interlocutore (Caron-Pargue e Caron, 1989). L'obiettivo di questo contributo è quello di studiare ed interpretare, da un punto di vista psicologico sociale, il contenuto del documento di rivendicazione dell'attentato delle Brigate Rosse (BR) a Marco Biagi, attraverso l'Analisi delle Corrispondenze (AC).

Il quadro teorico di riferimento adottato in questo studio è l'approccio proposto da Reicher, che integra la *Self-Categorization Theory* (SCT ; Turner, 1982) con elementi della psicologia retorico discorsiva (Reicher, 1996 ; Reicher e Hopkins, 1996). Molti ricercatori hanno studiato l'azione collettiva in funzione dell'identità collettiva, nella prospettiva della *Social Identity Theory* (Kelly, 1993 ; Kelly e Breinlinger, 1996 ; Klandermans e De Weerd, 2000 ; Simon e Klandermans, 2001), e vi sono diverse prove empiriche che mostrano come l'identificazione con il proprio gruppo sia di fatto il miglior predittore dell'azione collettiva (De Weerd e Klandermans, 1999 ; Kelly e Breinlinger, 1995 ; Kelly e Kelly, 1994 ; Klandermans, 2000). Tuttavia, ogni spiegazione della salienza e della definizione delle categorie collettive del sé, e del proprio ingroup, deve tener conto di come queste sono definite e argomentate nel linguaggio e nella retorica discorsiva. Inoltre, la definizione argomentativa di tali categorie svolge un ruolo fondamentale nella retorica di coloro che intendono invocare e formare un processo di mobilitazione di massa, come accade spesso, appunto, nei discorsi politici.

1.1. *Self-Categorization Theory e analisi discorsiva*

Com'è noto, la SCT sostiene che l'identità sociale costituisce il fondamento socio-cognitivo del comportamento di gruppo e che, quando le persone agiscono in termini di identità sociale, esse si percepiscono nei termini in cui è definito il proprio ingroup (Turner, 1982 ; Turner, Hogg, Oakes, Reicher e Wetherell, 1987). Alla base di tale processo vi è la categorizzazione della realtà sociale, che comporta un'accentuazione delle somiglianze intracategoriali e delle differenze intercategoriali. La categorizzazione saliente è quella in grado di spiegare in modo migliore le differenze e le somiglianze tra gli stimoli, mentre, secondo il principio del metacontrasto (Hogg e McGarty, 1990), la categoria saliente è quella che al contempo "minimizza le differenze intracategoriali e massimizza le differenze intercategoriali nell'ambito di uno schema di riferimento sociale" (Palmonari, 1995 : 420). In altri termini, questo processo implica un incremento della somiglianza percepita tra sé e i membri del proprio gruppo, una sorta di omogeneità intragruppo, e un incremento della dissomiglianza percepita tra sé e i membri dell'outgroup. Un membro di un gruppo si percepirà, quindi, come esempio prototipico dello stereotipo dell'ingroup, differenziandosi da tutti i membri dell'outgroup, percepiti a loro volta come esempi prototipici dell'outgroup. Tale "depersonalizzazione della percezione di sé" è il processo cognitivo alla base di vari fenomeni di gruppo, quali la stereotipizzazione sociale, la coesione di gruppo, l'etnocentrismo, e, come si è detto, l'azione collettiva.

Secondo Reicher, le modalità in cui le categorie sociali sono definite (cioè la loro inclusività, il loro contenuto e chi è considerato un tipico membro del gruppo) influenzano modi e forme delle mobilitazioni collettive (cioè la loro ampiezza, la loro direzione, gli obiettivi posti e la loro guida) (Reicher, 1996 ; Reicher e Hopkins, 1996). Particolare attenzione viene rivolta, quindi, al linguaggio con cui vengono discusse e trattate le questioni politiche e sociali, poiché il linguaggio è il dominio in cui vengono costruite e contestualizzate le definizioni delle categorie. Se è vero che la definizione delle categorie sociali può modulare la mobilitazione delle masse, allora il modo in cui gli eventi e gli attori sociali in essi implicati vengono caratterizzati, rappresenta uno dei modi attraverso cui influenzare l'azione collettiva. Chi si propone un obiettivo di mobilitazione collettiva, utilizzerà in maniera retorica un certo tipo di linguaggio e di argomentazioni al fine di definire identità e proposte politiche consonanti con l'orientamento del proprio gruppo/movimento, e dissonanti e in contrapposizione con quelli del gruppo/movimento avversario.

Queste strategie retorico-discorsive sono state analizzate ed illustrate da Reicher e Hopkins (1996) utilizzando come esempi i discorsi di Margaret Thatcher e da Neil Kinnock ai congressi dei rispettivi partiti, in occasione dello sciopero dei minatori inglesi avvenuto tra il 1984 e 1985. I risultati hanno evidenziato come le differenze tra i due politici possono essere fatti risalire ai diversi modi in cui essi intendono mobilitare la propria audience. La Thatcher costruisce una cornice di riferimento di "democrazia-contro-il-terrorismo", in cui la categoria inclusiva sono i conservatori e i "lavoratori-veri inglesi" contrari allo sciopero, mentre l'outgroup è costituito dagli "scioperanti-terroristi". Kinnock, invece, costruisce un quadro di riferimento di "Thatcher-contro-la-società" in cui la categoria inclusiva dell'ingroup è il popolo, la gente, i laburisti e i lavoratori inglesi a favore dello sciopero, mentre l'outgroup è costituito da Margaret Thatcher. Inoltre, proprio in funzione di questa diversa categorizzazione sociale, i valori evocati dai due politici, per far leva e mobilitare propri interlocutori, sono profondamente diversi. La "comunità nazionale" invocata dalla Thatcher è mobilitata sulla base di valori quali la risolutezza, il coraggio e l'ordine. Il "popolo" a cui si riferisce Kinnock, al contrario, è mobilitato sulla base di valori quali l'affetto, la compassione e la solidarietà.

1.2. *Obiettivi*

Alle 20.06 del 19 Marzo 2002, il professor Marco Biagi, consulente del ministero del Welfare, viene ucciso a Bologna da sei colpi di pistola davanti al portone di casa (per una ricostruzione della dinamica dell'attentato cf. Biacchessi, 2003). L'attentato viene rivendicato dalle "Brigate Rosse-partito comunista combattente" (BR-pcc). Come già accennato, lo scopo di questo studio è quello di interpretare, alla luce della SCT, la struttura del documento di rivendicazione dell'attentato a Marco Biagi utilizzando l'AC. Com'è noto, attraverso l'AC è possibile identificare un numero ristretto di dimensioni latenti, detti fattori, in grado di sintetizzare significativamente l'informazione contenuta nel lessico (Ercolani, Areni e Mannetti, 1990). Le unità lessicali ordinate lungo l'asse fattoriale possono essere concepite, in chiave linguistica, alla stregua di un sintagma (Bolasco, 1999), e possono assumere il valore di "sintagmi ideali", cioè frasi modali, teoriche, appartenenti all'intero corpus e non solo ad uno specifico enunciato. Tali sintagmi sono in grado di fornire "un paradigma, un modello di senso esistente all'interno del corpus" (Bolasco, 1999 : 232), che, vista la natura e gli intenti prefissi dal documento preso in esame nel presente studio, possono anche essere interpretati in un'ottica retorico-discorsiva.

2. Metodo

Materiale e analisi dei dati

Il materiale di studio è rappresentato dal testo di rivendicazione dell'attentato a Marco Biagi. Il documento venne reso accessibile al pubblico pochi giorni dopo l'attentato sul sito d'informazione www.caserta24ore.it, da cui è stato "scaricato" (cf. Benedetti, 2002). Il testo è composto in totale da 15444 parole, di cui 2705 diverse (17.5%).

La matrice di dati analizzata è quella *frammenti*forme* (Bolasco, 1999), in cui in riga si hanno i frammenti di testo del documento (segmenti di frasi), e in colonna si hanno le forme, cioè le parole. Su questo materiale è stata applicata l'AC su dati lessicali, utilizzando il pacchetto statistico Spad-t (Lebart, Morineau e Bécue, 1989), allo scopo di individuare le principali dimensioni latenti intorno alle quali si organizza il discorso politico veicolato dal documento.

3. Risultati

La prima fase dell'analisi è volta, come di consueto, alla *disambiguazione* del testo (Bolasco, 1999). Sull'insieme di parole con frequenza maggiore di 1 (in totale 13970 di cui 1232 composto da parole distinte, circa l'8.82%) si è provveduto a fare la procedura EQUIVALENCE per accorpare fra loro determinate forme semantiche. L'intervento sul testo, tuttavia, è stato ridotto al minimo, accorpando tra loro singolari e plurali (p.es. forza-forze ; classe-classi), e parole semanticamente simili (p.es. capitalismo-capitalistico-capitalistica ; imperialismo-imperialista-imperialisti-imperialiste). In alcuni sporadici casi, si sono accorparti sostantivi e verbi (p.es. governo, governare e governante ; trasformare, trasformazione e trasformazioni). Dopo la procedure di EQUIVALENCE, il numero di parole distinte è sceso a 1058 (7.57%). Le parole più frequentemente usate nel testo sono risultate essere : politica (f=275) ; rivoluzione (f=159) ; classe (f=144) ; imperialismo (f=127) ; stato (f=105) ; guerra (f=81) ; capitalismo (f=70) ; proletaria (f=68) ; processo (f=61) ; potere (f=52).

L'AC è stata effettuata sull'insieme di parole più significativamente presenti nel testo : nel nostro caso abbiamo considerato le parole con frequenza maggiore di 9 e con più di due lettere, per un totale di 156 parole distinte. L'AC ha fatto emergere quattro fattori latenti interpretabili. I primi due fattori risultano totalmente saturati da poche parole chiave, che fissano mezzi e fini politici dell'organizzazione. Il semiasse negativo del primo fattore è infatti composto dalla parola LOTTA e dalla parola ARMATA (tabella 1), che spiegano da sole il 97.2% dell'inerzia del semiasse. La lotta armata rappresenterebbe lo "strumento politico" adottato dall'organizzazione (ingroup).

Semiasse negativo						**Semiasse positivo**					
Parole	**c.a.**	**Coordinate**				**Parole**	**c.a.**	**Coordinate**			
		I	II	III	IV			I	II	III	IV
LOTTA	48.6%	-12.0	.0	.1	.0						
ARMATA	48.6%	-12.0	.0	.1	.0						
Inerzia spiegata	97.2%										

Tabella 1. Semiassi del primo fattore "Strumento politico dell'ingroup"

Il semiasse positivo del secondo fattore è, invece, composto dalle parole PARTITO, COMUNISTA, COMBATTENTE, COSTRUZIONE (tabella 2), che spiegano l'86.7% dell'inerzia

del semiasse. Le parole che compongono questo fattore sottolineano abbastanza chiaramente quale sia il "fine politico perseguito dall'organizzazione" (ingroup), e cioè la costruzione del partito comunista combattente.

Semiasse negativo						Semiasse positivo					
Parole	**c.a.**	**Coordinate**				**Parole**	**c.a.**	**Coordinate**			
		I	II	III	IV			I	II	III	IV
						PARTITO	29.5%	.12	7.0	1.0	-.69
						COMUNISTA	26.2%	.16	9.2	1.6	-1.2
						COMBATTENTE	25.3%	.17	9.1	1.7	-1.2
						COSTRUZIONE	5.7%	.09	1.8	-.43	.14
Inerzia spiegata							86.7%				

Tabella 2. Semiassi del secondo fattore "Fine politico dell'ingroup"

Mentre i primi due fattori sono costituiti da sintagmi tutto sommato ovvi e palesi, il terzo e il quarto fattore, al contrario, presentano delle contrapposizioni particolarmente interessanti. Il terzo fattore esprime l'obiettivo e la proposta politica dell'ingroup, contrapposto all'obiettivo politico ed economico perseguito dall'outgroup (tabella 3). Infatti, il semiasse negativo del terzo fattore può essere interpretato come un sintagma riconducibile all'obiettivo politico dell'organizzazione, che sarebbe la RIVOLUZIONE e l'instaurazione della DITTATURA PROLETARIA ad opera del PARTITO COMUNISTA combattente. Questo è in netta opposizione alla RIFORMA del MERCATO del LAVORO attualmente in corso nella società CAPITALISTA, e che rappresenta un NUOVO MODO di SFRUTTAMENTO e in cui gioca un RUOLO rilevante il SINDACATO (semiasse positivo).

Semiasse negativo						Semiasse positivo					
Parole	**c.a.**	**Coordinate**				**Parole**	**c.a.**	**Coordinate**			
		I	II	III	IV			I	II	III	IV
RIVOLUZIONE	5.0	.00	.04	-.97	.92	LAVORO	15.1	.19	-.45	3.6	2.3
MOVIMENTO	3.7	.36	-.04	-2.9	2.4	MERCATO	10.2	.24	-.65	4.6	3.0
DITTATURA	1.6	.14	-.64	-1.96	.18	RIFORMA	8.7	.23	-.72	4.3	2.4
PARTITO	1.4	.10	1.7	-1.5	.90	CAPITALISMO	5.1	.19	-.42	1.5	.13
PROLETARIA	1.3	.08	-.16	-.74	.47	CONDIZIONE	2.5	.15	-.31	1.3	.97
COMUNISTA	1.0	.02	1.1	-1.5	1.9	SINDACATO	2.3	.12	.45	2.3	-2.5
						LIVELLO	2.1	.08	.08	1.4	-.73
						MODO	1.6	-.02	.14	2.0	.78
						NUOVO	1.4	-.03	-.31	1.5	1.0
						SFRUTTAMENTO	1.4	-.04	-.12	1.8	1.4
						RUOLO	1.3	.12	1.1	1.5	-.19
Inerzia spiegata	14%						51.7%				

Tabella 3.
Semiassi del terzo fattore "obiettivo politico dell'ingroup vs. obiettivo economico outgroup"

Il quarto fattore sembra far riferimento agli "attori" contrapposti nella lotta politica in corso, determinando, in altre parole, l'ingroup e l'outgroup. Il semiasse negativo del quarto fattore è, infatti, composto da parole che individuano, definiscono e contestualizzano gli "avversari"

politici (l'outgroup), individuati nell'ESECUTIVO, agente dell'IMPERIALISMO, e nel SINDACATO, che attraverso la SUA AZIONE opera negli INTERESSI della BORGHESIA DOMINANTE. Il semiasse positivo, invece, definisce e contestualizza l'ingroup, ovvero le BRIGATE ROSSE, MOVIMENTO ANTIMPERIALISTA COMUNISTA COMBATTENTE.

Semiasse negativo						**Semiasse positivo**					
Parole	**c.a.**	**Coordinate**				**Parole**	**c.a.**	**Coordinate**			
		I	II	III	IV			I	II	III	IV
ESECUTIVO	10.4	.59	-.80	.58	-5.5	LAVORO	6.0	.19	-.45	3.6	2.3
IMPERIALISMO	8.7	.09	-.33	.11	-1.4	MERCATO	4.4	.24	-.65	4.6	3.0
SINDACATO	2.8	.12	.45	2.3	-2.5	ROSSE	3.0	-.65	.69	-.55	1.8
DOMINANTE	2.7	.08	-.23	-.04	-1.2	BRIGATE	2.6	-.25	.64	-.64	1.7
BORGHESIA	2.6	.17	-.38	.16	-1.2	MOVIMENTO	2.5	.36	-.04	-2.9	2.4
CATENA	2.3	.05	-.43	.12	-1.7	ANTIMPERIALISTA	1.7	.00	1.2	-1.2	2.2
AZIONE	2.1	.22	-.23	-.17	-1.4	COMUNISTA	1.7	.02	1.13	-1.5	1.9
INTERESSI	1.7	.11	-.39	-.07	-1.2	COMBATTENTE	1.4	-.04	1.3	-1.0	1.7
SUA	1.2	.02	-.23	-.33	-1.3						
Inerzia spiegata	34.5%						23.3%				

Tabella 4. Semiassi del quarto fattore "Ingroup vs. outgroup"

Incrociando il terzo e quarto fattore, si è ottenuto il piano in Figura 1, che rappresenta un'efficace sintesi dei risultati emersi. Com'è possibile notare, nel IV quadrante sono presenti parole che rimandano e definiscono la categoria inclusiva del sé collettivo, ovvero all'ingroup (PARTITO COMUNISTA COMBATTENTE MOVIMENTO ANTIMPERIALISTA), e, contemporaneamente, dei suoi obiettivi politici (RIVOLUZIONE e DITTATURA del PROLETARIATO). Al contrario, nel secondo quadrante viene definita la categoria esclusiva, ovvero l'outgroup (ESECUTIVO e SINDACATO), e i suoi fini politici (INTERESSI dell' IMPERIALISMO e della BORGESIA DOMINANTE). Il primo quadrante, invece, contiene parole che fanno riferimento al contesto in cui è inserita la retorica del documento (RIFORMA del MERCATO del LAVORO), e quindi, indirettamente, al quadro della lotta politica in corso (NUOVO MODO e CONDIZIONE di SRUTTAMENTO nel CAPITALISMO).

4. Discussione e conclusioni

In un'ottica psicologico sociale, la strategia retorica emersa dall'analisi del documento di rivendicazione assume particolare interesse se interpretato alla luce della *Self-Categorization Theory* (Reicher e Hopkins, 1996 ; Turner, 1982). Nella struttura del documento è, infatti, possibile rintracciare chiaramente la definizione di una categoria inclusiva del sé collettivo (ingroup), di una categoria esclusiva (outgroup) e di un contesto (frame). Il documento costruisce, infatti, una cornice di riferimento in cui la riforma del mercato del lavoro è il terreno in cui si esprime attualmente la lotta di classe, e in cui la categoria inclusiva sono le Brigate Rosse e il proletariato, mentre la categoria esclusiva è rappresentata dall'imperialismo e dalla borghesia dominante e i loro "agenti".

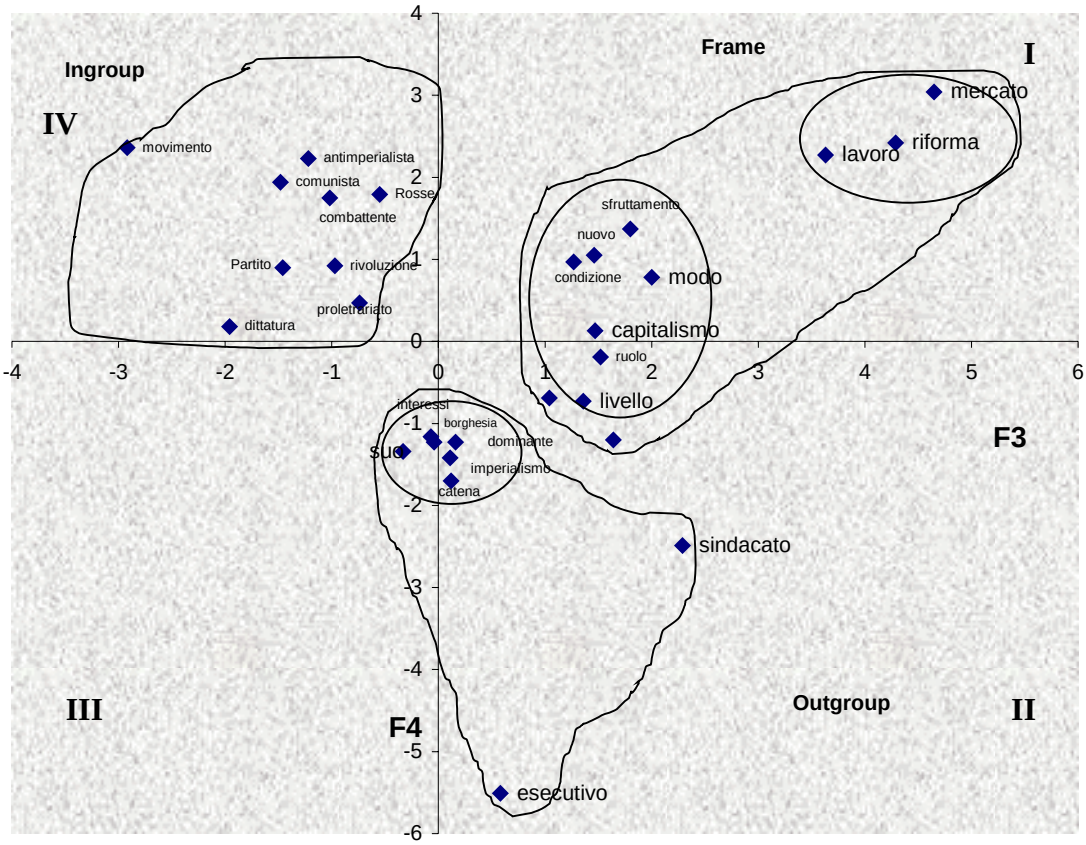


Figura 1. Piano fattoriale derivato dall'incrocio tra il III e IV fattore

Da una parte è quindi presente un ingroup fortemente caratterizzato nelle sue dimensioni organizzative e politiche, e cioè le Brigate Rosse-Partito Comunista Combattente. Dall'altra un outgroup rappresentato in due aspetti apparentemente contrapposti (ovvero il sindacato e l'esecutivo), ma considerate due facce della stessa medaglia, in quanto aventi obiettivi comuni, cioè la riforma del mercato del lavoro e gli interessi dell'imperialismo e della borghesia dominante. Inoltre, una porzione consistente della strategia retorica del documento di rivendicazione si impernia sull'affermazione della necessità, e della volontà, di perseguire la costruzione del partito comunista combattente attraverso la lotta armata, mentre si contrappongono nettamente le prospettive rivoluzionarie dell'ingroup a quelle riformiste dell' outgroup. Lo sfondo, il contesto, la contesa, la battaglia politica è rappresentata, invece, dalla riforma del mercato del lavoro in corso, considerata come un nuovo modo di sfruttamento capitalistico.

Bibliografia

Augoustinos M. (1998). Social Representations and Ideology : Towards the study of Ideological Representations. In Flick U. (Ed.), *The Psychology of the Social*. Cambridge University Press : 156-169.

Benedetti A. (2002). *Il linguaggio delle nuove Brigate Rosse*. Erga Edizioni.

Biacchessi D. (2003). *L'ultima bicicletta*. Mursia.

Billig M. (1991). *Ideology and Opinions.* Sage.

Bolasco S. (1999). *Analisi multidimensionale dei dati*. Carocci.

Caron-Pargue J. e Caron J. (1989). Processus psycholinqustiques et analyse des verbalitisations dans une tache cognitive. *Archives de Psycholgie,* vol. (57) : 3-32.

Catellani P. (1997). *Psicologia Politica*. Il Mulino.

De Weerd M. e Klandermans B. (1999).Group identification and political protest : Farmers' protest in the Netherlands. *European Journal of Social Psychology*, vol. (29/8) : 1073-1095.

Ercolani A. P., Areni A. e Mannetti L. (1990). *La ricerca in psicologia. Modelli di indagine e di analisi dei dati*. Carocci.

Hogg, M. e McGarty C. (1990). Self categorization and social identity. In Abrams D. e Hogg M. (Eds), *Social Identity Theory*. Harvester.

Kelly C. e Breinlinger S. (1995). Attitudes, intentions, and behavior : A study of women's participation in collective action. *Journal of Applied Social Psychology*, vol. (25/16) : 1430-1445.

Kelly C. e Breinlinger S. (1996). *The social psychology of collective action : Identity, injustice and gender*. Taylor & Francis.

Kelly C. (1993). Group identification, Intergroup Perceptions and Collective Action. In Stroebe W. e Hewstone H. (Eds.), *European Review of Social Psychology*, vol. (4) : 59-83.

Kelly C. e Kelly J. (1994). Who gets involved in collective action? Social psychological determinants of individual participation in trade unions. *Human Relations*, vol. (47) : 63-88.

Klandermans B. (2000). Identity and protest : How group identification helps to overcome collective action dilemmas. In Van Vugt M. e Snyder M. (Eds), *Cooperation in modern society : Promoting the welfare of communities, states and organizations.* Routledge : 162-183.

Klandermans B. e De Weerd M. (2000). Group identification and political protest. In Stryker S., Owens T.J. *et al*. (Eds), *Self, identity, and social movements. Social movements, protest, and contention*, vol. (13) : 68-90.

Lébart L., Morineau A. e Bécue M. (1989). *SPAD T (Système Portable pour l'Analise des Données Textuelles)*. Cisia.

Palmonari A. (1995). L'interazione nei gruppi. In Arcuri L. (Ed.), *Manuale di Psicologia Sociale*. Il Mulino : 365-424.

Reicher S. (1996). "The Battle of Westminster" : Developing the social identity model of crowd behaviour in order to explain the initiation and development of collective conflict. *European Journal of Social Psychology*, vol. (26) : 115-134.

Reicher S., e Hopkins N. (1996). Self-category constructions in political rhetoric : An analysis of Thatcher's and Kinnock's speeches concerning the British miners' strike (1984-5). *European Journal of Social Psychology*, vol. (26) : 353-371.

Sensales G., Chirumbolo A., Areni A. e Kosic A. (2002). Libere associazioni e rappresentazioni della "politica" di studenti universitari de "La Sapienza". Analisi del ruolo giocato da costrutti di personalità nella ricostruzione discorsiva. In : A. Morin e P. Sébillot (Eds.), *Sixth International Conference on Textual Data Statistical Analysis*, vol. (2) : 713-722

Sensales G., Chirumbolo A. e Areni A. (2002). *Giovani e Politica*. Kappa.

Sensales S., Chirumbolo A., Areni A. e Bettini F. (in stampa). Representations of "politics" : a pilot survey among students of 'La Sapienza' University of Rome. *Ricerche di Psicologia*.

Simon B. e Klandermans B. (2001). Politicized collective identity : A social psychological analysis. *American Psychologist*, vol. (56/4) : 319-331.

Turner J.C. (1982). Towards a cognitive redefinition of the social group. In Tajfel H. (Ed.), *Social identity and intergroup relations*. Cambridge University Press.

Turner J.C., Hogg M., Oakes P.J., Reicher S.D. e Wetherell M. (1987). *Rediscovering the social group*. Blackwell.

Van Dijk T. (1998). *Ideology*. Sage Pubblications.

Comparative study of statistical word sense discrimination techniques

Marie-Catherine de Marneffe, Pierre Dupont

Computing Science Department, INGI – UCL
1348 Louvain-la-Neuve – Belgium
{mcdm, pdupont}@info.ucl.ac.be

Abstract

Word sense discrimination aims at automatically determining which instances of an ambiguous word share the same sense. A fully unsupervised technique based on a vector representation of word senses was proposed by Schütze (Schutze, 1998). While the original model was assumed to be Gaussian, practical results were only reported for an approximated model making hard decisions between sense clusters. We show in the present study that a real Gaussian model provides a significant accuracy improvement while remaining fully tractable. An alternative discrete naïve Bayes model was presented in Manning and Schütze (1999). We propose here a description of both models in a unified statistical formalism in order to stress the similarities and differences between both approaches. Several practical experiments are conducted on the New York Times News 1997 corpus. They illustrate the respective advantages of various approaches trading off discrimination accuracy and computation time. We also show the interest of a global selection of content words to characterize the context of an ambiguous instance in the naïve Bayes model.

Keywords: word sense disambiguation, discrimination techniques, Naïve Bayes, K-means, expectation-maximization algorithm.

1. Introduction

The purpose of automatic word sense disambiguation is to determine the exact sense of an instance of an ambiguous word according to its particular use. Disambiguation can be useful in principle in any linguistic application where word sense matters such as automatic translation, text categorization, speech understanding, *etc*.

1.1. *Word sense disambiguation techniques*

Word sense disambiguation techniques can be divided into three broad categories: supervised techniques, dictionary-based (or thesaurus-based) and unsupervised techniques. All these techniques use the possible *senses* of the ambiguous word, the *contexts* of the instances of the ambiguous word and some sense *informants*.

Supervised techniques require a semantically tagged corpus, which serves as training corpus, in which each ambiguous instance w is correctly labeled with a semantic tag. The possible *senses* are defined by the set of semantic tags present in the corpus. The *contexts* consist of a window around instances of w, possibly limited to the syntactic group of w, and *informants* are the words belonging to those context windows. For example, Gale *et al.* use a *naïve Bayes classifier* to disambiguate words: the training corpus enables to assign to each informant the probability that it induces a sense (Gale *et al.*, 1992). Brown *et al.* propose an *information theoretic approach* which gives a sense to an ambiguous word as used for translation (Brown *et al.*, 1991). This technique determines the different values of the best informant. For instance, "prendre la

voiture" in French is translated in English by "to *take* the car" and "prendre une décision" by "to *make* a decision". Here the informant is the verb object. Once the informant and its values have been found, an algorithm based on *mutual information* is applied to determine which informant value induces a specific translation. Yarowsky uses an alternative approach based on *decision lists* (Yarowsky, 1994). An ordered list of informants is built from the training corpus, the most salient informants appearing first in the list. Each informant is associated to one sense. Disambiguation of a new instance is based on the first informant in the decision list which appears in the instance context. Ng and Lee (1996) propose an *exemplar-based approach*. The sense of an ambiguous word is determined by the instance which appears in the most similar phrase found in the training corpus. Several approaches have been compared in the *Senseval* project, a systematic evaluation of supervised techniques for word sense disambiguation (Kilgarriff, 1998; Kilgarriff and Rosenzweig, 2000).

Dictionary-based techniques work similarly as the supervised techniques but use a raw (i.e. untagged) corpus. A dictionary or a thesaurus is an additional knowledge source to define senses. In Lesk's algorithm (Lesk, 1986) the sense of an ambiguous word instance w is determined by the dictionary definition having the largest number of words in common with w context. Yarowsky proposes another approach based on the semantic categorization of the *Roget's International Thesaurus* (Yarowsky, 1992). The informants are words that often occur in the context of a semantic category of the *Roget's*.

We study here the third category of disambiguation techniques which are fully unsupervised. In such case, a particular sense cannot be assigned to an ambiguous instance. Here the problem is to automatically determine which instances can be clustered as sharing the same sense, the sense labels being arbitrary. This task can be performed through unsupervised clustering of word contexts which represent the unknown senses. Dictionary-based techniques are sometimes also referred to as unsupervised techniques since they do not require a semantically tagged corpus. To make this distinction clear, we refer to fully unsupervised disambiguation as word sense *discrimination*.

1.2. *A comparative study of statistical word sense discrimination*

Schütze's technique for word sense discrimination is based on a vector representation of the word contexts (Schütze, 1998). Unsupervised clustering of word senses is performed in vector space. Assuming a Gaussian distribution for each cluster, this model can be estimated with the Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977). The practical results reported by Schütze are actually based on a simplified model estimated with the K-means algorithm (Duda and Hart, 1973). This simplification was introduced for computational efficiency reasons. The first objective of this paper is to study the impact of this simplification: is there a performance gain when a real Gaussian model is estimated and, if so, at which additional computational cost?

Schütze's approach is also mentioned in Manning and Schütze (1999) but the probabilistic model used in this case is a discrete model of word contexts with naïve Bayes assumption. This model can also be estimated with the EM algorithm but it differs from the vector model. The discrete versus continuous nature of these models is one evidence of this distinction. The second objective of this work is to clarify this distinction and to study the relative performances of both approaches.

Statistical word sense discrimination is formally presented in section 2. Discrimination based on a discrete modeling of word contexts is described in section 3. The two variants (EM or K-means estimation) of the vector model are presented in section 4. Several experiments have been

performed on the *New York Times News*. Section 5 details the corpus and our experimental protocol. Comparative results are presented in section 6.

2. Statistical word sense discrimination

In the sequel we use the following notations:
- w denotes an ambiguous word,
- $s_1, \ldots, s_K$ denote the K possible senses[1] of w,
- $c_1, \ldots, c_I$ denote the contexts of the I instances of w in a training corpus,
- $v_1, \ldots, v_J$ denote J possible informants.

Following Bayes decision theory (Duda *et al.*, 2001), word sense discrimination can be formulated as computing the sense $\hat{k}$ which maximizes the *posterior probability* $P(s_k|c)$ of sense s_k given the observed context c:

$$\hat{k} = \underset{k}{\operatorname{argmax}}\, P(s_k|c) = \underset{k}{\operatorname{argmax}}\, P(c|s_k)P(s_k), \tag{1}$$

where $P(c|s_k)$ is the *likelihood* of context c given the sense s_k, and $P(s_k)$ denotes the *prior* probability of sense s_k.

A discrimination model defines how the context likelihoods and prior probabilities can be computed from a set of parameters Θ. These parameters are estimated from an unlabeled training corpus, generally depending on some informants. How these parameters are estimated and which are the informants depend on the particular approach, as detailed in the following sections.

3. Naïve Bayes word sense discrimination

Given a context c_i of w, that is a window around an instance of w in the training corpus, the informants are the context words v_j. These are the content words (as opposed to stop words[2]) belonging to c_i. The context likelihood is defined as a joint probability:

$$P(c_i|s_k) = P(\{v_j \in c_i\}|s_k).$$

According to the naïve Bayes assumption, the context words are assumed to be independent[3]. In other words, the joint probability can be rewritten as

$$P(\{v_j \in c_i\}|s_k) = \prod_{v_j \in c_i} P(v_j|s_k).$$

The set of parameters Θ for each word w consists of the $J.K$ probabilities $P(v_j|s_k)$ and the K priors $P(s_k)$. These parameters can be estimated so as to maximize the likelihood of a training corpus. As the corpus is untagged, this is a problem of incomplete data which can be solved by the EM algorithm (Dempster *et al.*, 1977).

[1] In word sense discrimination the $s_1, \ldots, s_K$ labels are arbitrary but the number K of possible senses must be decided. Automatic determination of an optimal K could also be considered.

[2] As detailed in section 5.1, stop words are conjunctions, prepositions, articles and other words, which appear often in documents yet alone may contain little meaning.

[3] This assumption is strongly arguable from a linguistic viewpoint but drastically reduces the number of parameters to be estimated and works surprisingly well in practice. Note also that, in the context of Bayes decision theory, this assumption can be reformulated in a more acceptable way as: $\operatorname{argmax}_k P(s_k)P(\{v_j \in c_i\}|s_k) = \operatorname{argmax}_k P(s_k) \prod_{v_j \in c_i} P(v_j|s_k)$.

The EM algorithm is an iterative procedure which, starting from an initial guess Θ^0 of the parameter values, recomputes in each iteration the parameter estimates so as to increase the data likelihood or, equivalently, its log-likelihood. The log-likelihood LL of the I contexts observed in the training corpus is defined as follows[4]:

$$\begin{aligned}\mathrm{LL}(\{c_1,\ldots,c_I\}|\Theta) &= \log\prod_{i=1}^{I}P(c_i)=\sum_{i=1}^{I}\log P(c_i)\\ &= \sum_{i=1}^{I}\log\sum_{k=1}^{K}P(c_i|s_k)P(s_k)\\ &= \sum_{i=1}^{I}\log\sum_{k=1}^{K}P(s_k)\prod_{v_j\in c_i}P(v_j|s_k). \qquad (2)\end{aligned}$$

In practice, the $P(v_j|s_k)$ are randomly initialized while satisfying the constraints: $\sum_{j=1}^{J}P(v_j|s_k)=1$, $1\leqslant k\leqslant K$, and uniform priors are assumed: $P(s_k)=\frac{1}{K}$. The two steps of the EM algorithm are then computed iteratively as long as the log-likelihood increases.

E-step: Compute h_{ik}, an estimate of the posterior probability that sense s_k generated c_i:

$$h_{ik}=\frac{P(s_k)P(c_i|s_k)}{\sum_{l=1}^{K}P(s_l)P(c_i|s_l)}=\frac{P(s_k)\prod_{v_j\in c_i}P(v_j|s_k)}{\sum_{l=1}^{K}\left(P(s_l)\prod_{v_j\in c_i}P(v_j|s_l)\right)}.$$

M-step: Re-estimate $P(v_j|s_k)$ and $P(s_k)$ so as to maximize the likelihood:

$$P(v_j|s_k)=\frac{\sum_{\{c_i:v_j\in c_i\}}h_{ik}}{\sum_{j=1}^{J}\sum_{\{c_i:v_j\in c_i\}}h_{ik}},$$

where $\sum_{\{c_i:v_j\in c_i\}}h_{ik}$ sums over all contexts c_i in which v_j occurs.

$$P(s_k)=\frac{\sum_{i=1}^{I}h_{ik}}{\sum_{k=1}^{K}\sum_{i=1}^{I}h_{ik}}=\frac{\sum_{i=1}^{I}h_{ik}}{I}.$$

Once the parameters of the model have been estimated on the training corpus, the sense of a new instance of w can be assigned based on its context c. The final decision rule is:

$$\hat{k}=\operatorname*{argmax}_k P(s_k|c)=\operatorname*{argmax}_k P(s_k)\prod_{v_j\in c}P(v_j|s_k)=\operatorname*{argmax}_k\ \log P(s_k)+\sum_{v_j\in c}\log P(v_j|s_k). \qquad (3)$$

4. Vector-based word sense discrimination

In the original Schütze's approach, words, contexts and senses are represented in a high-dimensional real-valued vector space (Schütze, 1998). Word vectors, context vectors and senses (clusters of context vectors) are presented in section 4.1. The probabilistic model and two variants of the estimation algorithm are described in sections 4.2. and 4.3.

[4] We follow here the presentation in Manning and Schütze (1999) but we use the corrected formulas as described in http://nlp.stanford.edu/fsnlp/errata.html.

4.1. *Vector representation of senses*

Word vector

A word w can be represented by a vector in which each component corresponds to a word v occurring in the corpus. The vector components represent frequencies of *co-occurrence*: the component associated with word v is the number of times that v occurs as a neighbor of w in the corpus. A neighbor is a content word occurring in a context window centered on w. These content words are the informants in this approach. For instance, if the words *legal* and *clothes* appear respectively 300 and 75 times in context windows of the word *judge*, the vector for *judge* can be represented as follows.

$$\begin{array}{r} \\ legal \\ \\ clothes \\ \\ \end{array} \begin{array}{c} judge \\ \left[\begin{array}{c} \dots \\ 300 \\ \dots \\ 75 \\ \dots \end{array}\right] \end{array}$$

Schütze examines two different ways to choose the vector dimensions: a local selection which focuses on words occurring as neighbors of the ambiguous word and ignores the rest of the corpus; a global selection which chooses the 2,000 most frequent words in the entire corpus. Moreover *word vectors* are computed only for the 20,000 most frequent words of the corpus. A 2,000-by-20,000 co-occurrence matrix can thus be derived from the corpus. To compute the most frequent words of the corpus, stop words are excluded. The best results were obtained using global selection.

Context vectors and senses

The context of an instance w is represented by a vector $\vec{x}$ obtained as the weighted sum of the *word vectors* of w neighbors (second-order co-occurrence). Given the *word vectors* $\vec{v_j}$, the *context vector* $\vec{x}$ is defined as

$$\vec{x} = \sum_{v_j \in c} a_j \vec{v_j}.$$

The weight a_j of vector $\vec{v_j}$ depends on the inverse document frequency (idf), a measure of its discriminative capability:

$$a_j = -\log \frac{d_j}{D},$$

where D denotes the number of documents in the corpus and d_j the number of documents in which v_j occurs (see section 5.1. for additional details on the corpus).

Similar *context vectors* can be seen as forming clusters in vector space. Each cluster represents one sense of an ambiguous word and can be characterized by its mean and covariance matrix. A new instance w is represented by its *context vector*. The sense of w is then assigned to the most similar cluster. Two different ways of defining the clusters are described in sections 4.2. and 4.3.

4.2. *Gaussian modeling of context clusters*

Context clusters are assumed to follow a Gaussian distribution. The whole model is a mixture of K Gaussian components, with one mixture component for each sense. Let $\vec{x_1}, \dots, \vec{x_I}$ denote the context vectors ($\vec{x_i} \in \mathbb{R}^d$ is the vector associated to context c_i). $\omega_1, \dots, \omega_K$ are the K

components. Each component ω_k is characterized by some parameters: the prior probability $P(s_k)$, the mean vector $\vec{\mu_k}$ and the covariance matrix Σ_k.

Starting from an initial guess of the parameter values Θ^0, these parameters are reestimated with the EM algorithm so as to maximize the training data likelihood. The initialization procedure typically follows from a hard clustering of the context vectors as detailed in section 4.3. Such clustering defines a first estimate of the K mean vectors. The context vectors are then assigned to their closest mean and the cluster covariance matrices can be computed. The initial prior of cluster ω_k is defined as $P(s_k) = \frac{i_k}{\sum_{k=1}^{K} i_k} = \frac{i_k}{I}$ where i_k is the number of vectors assigned to cluster ω_k and I is the total number of context vectors.

The log-likelihood of the I contexts observed in the training corpus is defined as

$$\mathrm{LL}(\{\vec{x_1}, \ldots, \vec{x_I}\}|\Theta) = \log \prod_{i=1}^{I} P(\vec{x_i}) = \sum_{i=1}^{I} \log \sum_{k=1}^{K} P(s_k) f_k(\vec{x_i}), \tag{4}$$

where $f_k(\vec{x_i})$ denotes the value of the Gaussian density in $\vec{x_i}$

$$f_k(\vec{x_i}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} \exp\left[-\frac{1}{2}(\vec{x_i} - \vec{\mu_k})^T \Sigma_k^{-1} (\vec{x_i} - \vec{\mu_k})\right].$$

The two steps of the EM algorithm are then computed iteratively as long as the log-likelihood increases.

E-step: From the parameter values at iteration r, compute h_{ik}, the posterior probability that ω_k generated $\vec{x_i}$:

$$h_{ik} = \frac{P(s_k) f_k(\vec{x_i})}{\sum_{l=1}^{K} P(s_l) f_l(\vec{x_i})}.$$

M-step: Re-estimate the parameters at iteration $r+1$ so as to maximize the likelihood:

$$\begin{aligned}
\vec{\mu_k}^{r+1} &= \frac{\sum_{i=1}^{I} h_{ik} \vec{x_i}}{\sum_{i=1}^{I} h_{ik}}, \\
\Sigma_k^{r+1} &= \frac{\sum_{i=1}^{I} h_{ik} (\vec{x_i} - \vec{\mu_k}^r)^T (\vec{x_i} - \vec{\mu_k}^r)}{\sum_{i=1}^{I} h_{ik}}, \\
P^{r+1}(s_k) &= \frac{\sum_{i=1}^{I} h_{ik}}{\sum_{l=1}^{K} \sum_{i=1}^{I} h_{il}} = \frac{\sum_{i=1}^{I} h_{ik}}{I}.
\end{aligned}$$

Once the parameters have been estimated on the training corpus, the sense of a new instance of w can be assigned from the vector $\vec{x}$ associated to its context c. The final decision rule is:

$$\hat{k} = \operatorname*{argmax}_k P(s_k) P(c|s_k) = \operatorname*{argmax}_k P(s_k) f_k(\vec{x}). \tag{5}$$

4.3. *Hard clustering of context vectors*

The probabilistic model described in section 4.2. defines a Gaussian mixture of K components. Any context vector $\vec{x}$ can be seen as being generated by all K components. This approach is sometimes called soft-clustering since a vector is not deterministically assigned to a particular cluster (i.e. a mixture component). An alternative approach is hard clustering where $\vec{x}$ is

assigned to its closest cluster mean according to the euclidean distance in vector space. Hard clustering can either be used as initialization before reestimation of a Gaussian model or as a sense discrimination technique as such.

Hard clustering can be performed in two steps with group-average agglomerative clustering (GAAC) and K-means (Schütze, 1998). GAAC is a bottom-up hierarchical clustering algorithm. Starting from a randomly selected subset of the I context vectors, GAAC iteratively agglomerates vectors into K clusters by merging most similar vectors first. The similarity measure used is the cosine:

$$sim(\vec{x}, \vec{y}) = cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|},$$

which simply amounts to the dot product for normalized vectors. The computational complexity of GAAC is $\mathcal{O}(n^2)$ where n denotes the size of the initial vector subset. In practice, a subset of $\sqrt{I}$ vectors can be selected. This allows to compute reasonably good cluster means in $\mathcal{O}(I)$.

The K cluster means serve as initialization for the K-means algorithm which runs in $\mathcal{O}(I)$ (Duda *et al.*, 2001). The I context vectors are first assigned to their closest means. Cluster means $\vec{\mu_k}$ are then recomputed. This process is iterated as long as the $\vec{\mu_k}$ vectors change.

Once the parameters have been estimated on the training corpus, the sense of a new instance of w can be assigned from the vector $\vec{x}$ associated to its context c by finding its closest mean. The final decision rule is:

$$\hat{k} = \underset{k}{\operatorname{argmin}} \parallel \vec{x} - \vec{\mu_k} \parallel . \tag{6}$$

Equation (6) is equivalent to equation (5) provided the K senses are assumed equally likely ($P(s_k) = \frac{1}{K}$) and a common covariance matrix is assumed for all K mixture components. This property illustrates that the model presented here is a simplified version of the Gaussian mixture model presented in section 4.2. Note however that the K estimated means are not necessarily the same in both models.

5. Experimental Assessment

Section 5.1. describes the corpus used in our experiments. The role of pseudowords and how they are used is described in section 5.2. Other details of the experimental protocol are presented in section 5.3. As detailed in the sequel, we follow here as much as possible Schütze's choices in parameter setting for comparison purposes.

5.1. *Corpus and stop list*

The available corpus selected for our experiments is the *New York Times News* of 1997. The training set comes from the first six months issues (January 1997 till June 1997). It contains 74,847,796 word tokens ($\sim$ 500 megabytes). There are 485,936 different words in this set. The words are not stemmed: singular and plural forms of a same word count for two different words and each form of a same verb counts for a different word.

This corpus is divided into *documents*. Each document corresponds to an article in the newspaper. The training set is made of 116,010 documents. The mean number of words per document is 645 with a standard deviation of 392.

The test set is extracted from the first 17 days of December 1997. The test set contains 7,857,354 word tokens ($\sim$ 50 megabytes) among which 135,502 different words. The mean number of words per document is 621 and the standard deviation is 397.

The proportion between training and test set have been chosen so as to represent a similar amount of data[5] as in Schütze's experiments (Schütze, 1998). The same context window size (50 tokens) have been chosen as well. The context window of an instance of w is made of up to 25 tokens on the left and 25 tokens on the right of w. A context window never crosses the limit of a document and only content words are considered inside it. Content words are defined as any word not belonging to a stop list. Our stop list is made of 574 stop words as defined in http://lingo.lancs.ac.uk/devotedto/corpora/software.htm[6]. Stop words are conjunctions, prepositions, articles and other words, which appear often in documents yet alone may contain little meaning.

5.2. *Pseudowords*

Ambiguous word	Sense	Distribution	**Pseudoword**	Senses	Training	Test
accident	chance	14%	banana-moon	banana	452	39
	crash	86%		moon	2,452	263
				Total	2,904	302
motion	physical movement	39%	animal-river	animal	2,389	219
	proposal for action	61%		river	6,104	362
				Total	8,493	581
train	to teach	30%	rely-illustration	rely	1,669	149
	line of railroad cars	70%		illustration	3,541	334
				Total	5,210	483
interest	feeling of special attention	31%	data-school	data	9,154	1,032
	charge on borrowed money	69%		school	29,095	2,468
				Total	38,249	3,500
suit	set of garments	12%	railway-admission	railway	550	27
	action or process in a court	88%		admission	1,974	189
				Total	2,524	216

Table 1. Pseudoword frequencies.

In order to test the performance of sense discrimination algorithms on naturally ambiguous words, a large number of instances have to be disambiguated by hand. As this is a time-consuming task, it is convenient to generate artificially ambiguous words: *pseudowords*. A pseudoword is the concatenation of two or more natural words.

Discrimination of pseudowords does not exactly reflect the discrimination task of real ambiguous words but precautions can be taken so as to best reflect a natural case (Gaustad, 2001). For example, the real ambiguous word *accident* has two main senses: *crash* and *chance*. A hundred instances of *accident* were manually tagged to determine its sense distribution. The corpus is then searched for two unambiguous words having a frequency of occurrence roughly fitting the ambiguous word sense distribution. In the case of *accident*, the unambiguous words *banana* and *moon* satisfy this requirement. All instances of *banana* and *moon* in the training corpus are then replaced by the pseudoword *banana-moon*. Table 1 gives the pseudowords built for five natural ambiguous words (with their respective sense distribution) and their frequencies of occurrence in the training and test sets.

[5] Schütze used the *New York Times News* of 1989-90. His training and test sets contain respectively 60.5 million word tokens and 5.4 million word tokens.

[6] Our stop list is the union of 4 stop lists found under the reference *Function Words/Stop Lists for English*.

5.3. *Experimental protocol*

All discrimination models include some random initialization before reestimation. In the naïve Bayes discrimination model (section 3.), the likelihoods $P(v_j|s_k)$ are initialized at random. In the hard clustering discrimination model (section 4.3) the K mean vectors derive from a randomly selected subset of the I context vectors. The result of the estimated K means is also used to initialize a Gaussian model (section 4.2.). As the EM algorithm is only guaranteed to find a local optimum of the likelihood, its performance depends indirectly on this initialization. Hence all experiments are repeated 10 times while varying the random seeds. Averaged results over these 10 independent runs are reported in section 6. In all experiments so far, the value of K is equal to 2 (binary sense discrimination).

The implementation used for the Gaussian model assumes a diagonal covariance matrix for each cluster. Possible correlations between the components of the context vectors are ignored but the number of parameters to be estimated for each ambiguous word is reduced to $K(1 + 2d)$, where d denotes the dimension of the vector space. In all cases, the result of the hard clustering techniques (the K means representing Kd parameters) was used to initialize the Gaussian model. Hence we were able to check whether the Gaussian model further improves the performance obtained with hard clustering. In all tests of the vector model a global selection of the vector dimensions was chosen (see section 4.1.).

Pseudoword	Occurrences (I)	Context words (J)	Parameters
banana-moon	2,904	11,449	22,900
animal-river	8,493	27,413	54,828
rely-illustration	5,210	17,763	35,528
data-school	38,249	56,008	112,018
railway-admission	2,524	11,593	23,186
Average	11,512	24,845	49,692

Table 2. Number of occurrences and informants in the local naïve Bayes approach.

In the naïve Bayes approach (section 3), the J informants to discriminate the senses of an ambiguous instance w are the J content words belonging to context windows around w in the training corpus. This implies that the number J and identity of informants depend on the word w. We refer to this approach as *local naïve Bayes*. The total number of parameters of the local model is $K(1 + J)$. Table 2 reports the number of informants for each pseudoword and the corresponding number of parameters.

An alternative approach is to consider the same set of informants for all ambiguous words. In this case the 20,000 most frequent content words in the training corpus are considered. The number of parameters (40,002) does no longer depend on the word to disambiguate. We refer to this approach as *global naïve Bayes*.

6. Results

Table 3 gives the discrimination results for the pseudowords considered in these experiments. The first two measures (S1, S2) for each pseudoword give the percentage of correct senses for each of the two words making the pseudoword. As the sense labels are arbitrary in a sense discrimination experiment, the most frequent sense (S1) is considered to be attributed to the most frequent word in the training (e.g. *moon* for the *banana-moon* pseudoword).The accuracy gives the total proportion of correctly labeled instances for both senses. In each case, the mean (μ) and standard deviation (σ) of the performances obtained over 10 independent runs are reported.

The last line reports the average accuracy (mean and standard deviation) obtained for the five pseudowords.

		Naïve Bayes				Vector Model			
Pseudowords		**Local**		**Global**		**K-Means**		**Gaussian**	
		μ	σ	μ	σ	μ	σ	μ	σ
banana-moon	S1	56.9	2.4	58.1	2.0	68.4	0.0	100.0	0.0
	S2	34.4	17.7	27.2	12.3	0.0	0.0	0.0	0.0
	accuracy	54.0	2.8	54.1	1.5	59.6	0.0	87.1	0.0
animal-river	S1	66.4	14.2	76.3	15.7	73.8	13.1	84.1	0.4
	S2	34.4	17.8	29.5	14.6	20.8	7.0	7.5	0.2
	accuracy	54.3	7.0	58.7	6.4	53.8	5.5	55.2	0.2
illustration-rely	S1	88.7	5.9	87.9	7.2	81.9	5.9	99.9	0.2
	S2	35.2	11.2	31.4	10.3	55.5	39.7	0.5	0.7
	accuracy	72.2	3.9	70.5	5.9	73.8	8.8	69.2	0.1
data-school	S1	78.2	16.0	92.3	9.9	83.1	0.9	93.5	0.2
	S2	40.0	34.5	59.0	16.3	58.8	17.0	44.4	15.0
	accuracy	66.9	14.7	82.5	4.5	75.9	4.4	79.0	0.3
railway-admission	S1	76.8	12.7	79.3	9.9	56.2	15.4	99.5	0.2
	S2	32.9	23.5	48.5	21.1	67.0	22.3	13.3	4.7
	accuracy	71.3	11.2	75.5	6.7	57.5	10.7	88.8	0.4
Average accuracy		63.7	7.9	68.3	5.0	64.1	5.9	75.9	0.2

Table 3. Discrimination results.

The average accuracy illustrates that the two reference models proposed respectively in Manning and Schütze (1999) and Schütze (1998), namely the local naïve Bayes and K-means vector model, perform roughly as well. Note however that for a given pseudoword these approaches can give significantly different results. For instance, the K-means vector model wrongly attributes all instances of *banana* to the *moon* cluster. Moreover this result is not affected by the random initialization as it does not change over the 10 independent runs ($\sigma = 0$). In contrast, the local naïve Bayes model splits test instances between the 2 senses.

The global naïve Bayes model slightly improves over the reference models. The Gaussian vector model performs significantly better on average than the reference models. Moreover the variance of the results is also decreased showing that this approach is less sensitive to a particular initialization. Note that the Gaussian model tends to favor the majority sense in several cases.

Table 4 summarizes the computational cost for estimating[7] the discrimination models and the number of estimated parameters[8] in each case. As the number of occurrences of the pseudowords varies in the training set (see table 2), the reported CPU times[9] correspond to the estimated times for processing 3,000 occurrences in all cases. This analysis can probably be refined with a detailed profiling and optimization of the programming code but it illustrates already the tractability of all approaches considered so far. The Gaussian model offers the best accuracy and is parsimonious as it has 5 times less parameters than the global naïve Bayes model. The average number of iterations required to converge is reported in the last column.

[7] The figure reported corresponds to the time for estimating the discrimination models from precomputed context vectors or context windows. Hence this time does not include the preprocessing of the corpus to extract the contexts and filter out the stop words.

[8] For the local naïve Bayes model, the average number of parameters has been reported (see table 2).

[9] The CPU times are measured on a laptop with a 600 MHz processor and 384 Mb of RAM. All estimation programs are written in C.

Method	Accuracy	CPU Time (sec)	Number of parameters	Iterations
Naïve Bayes (local)	63.7	.4	49,692	12
Naïve Bayes (global)	68.3	.5	40,002	17
Vector Model (K-means)	64.1	10.5	4,000	10
Vector Model (Gaussian)	75.9	36.7	8,002	6

Table 4. Accuracy/CPU Time trade-off.

7. Conclusion and future work

We compared in this work several word sense discrimination techniques. The vector model proposed by Schütze (1998) can significantly be improved when a real Gaussian model is estimated instead of its hard clustering approximation. This performance gain is obtained with an additional computational cost but the estimation procedure remains very efficient in all cases. The Gaussian model tested here includes a diagonal covariance matrix for each sense. We could also consider a full covariance matrix but this would significantly increase the number of parameters and the computation time. This option will be evaluated in further experiments.

The naïve Bayes model described in Manning and Schütze (1999) has also been implemented and its average performance is comparable with the hard clustering approach. Our experiments demonstrate that a performance gain is obtained when the same context informants are used for all pseudowords. This global approach has the advantage of a common set of parameters for all ambiguous words which are more reliably estimated over the whole training corpus.

Our results are not fully comparable with Schütze's experiments even though we followed the same experimental protocol, as closely as possible. The first reason is that the used corpora differ (New York Times News 97 versus 89-90) but a similar amount of data was used. The stop lists differ (574 words versus 930 words). The pseudowords were also built in a slightly different way. We argue that our pseudoword design better reflects the discrimination task for naturally ambiguous words while not requiring time consuming labeling of the corpus. Schütze also demonstrated the advantage of reducing the vector space dimension with *Singular Value Decomposition* (SVD) (Berry, 1992). Including SVD in the vector model is our very next task.

Several additional options and extensions will be considered in the future. In particular we will study:

- the influence of the context window size (currently 50 words around the ambiguous instance); we expect that this size can be significantly reduced,
- the influence of stemming and the definition of the stop list,
- the number K of senses considered (currently only binary sense discrimination is considered), and the automatic determination of an optimal K value,
- the dimension of the original vector space and the final space dimension after singular value decomposition,
- smoothing techniques to improve estimates of the global naïve Bayes model.

References

Berry M.W. (1992). Large-scale sparse singular value decomposition. *The International Journal of Supercomputer Applications*, vol. (6/1): 13-49.

Brown P., Della Pietra S., Della Pietra V. and Mercer R. (1991). Word-sense disambiguation using statistical method. In *Proceedings of ACL*, vol. (29): 139-145.

Dempster A., Laird N. and Rubin D. (1977). Maximum Likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society Ser. B (methodological)*, vol. (39): 1-38.

Duda R. and Hart P. (1973). *Pattern Classification and Scene Analysis*. Wiley.

Duda R., Hart P. and Stork D. (2001). *Pattern Classification*. Wiley.

Gale W., Church K. and Yarowsky D. (1992). A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, vol. (26): 415-439.

Gaustad T. (2001). Statistical Corpus-Based Word Sense Disambiguation: Pseudowords vs Real Ambiguous Words. In *Companion Volume to the Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL/EACL 2001) — Proceedings of the Student Research Workshop*.

Kilgarriff A. (1998). Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs. *Computer Speech and Language*, vol. (12/4): 453-472.

Kilgarriff A. and Rosenzweig J. (2000). English Senseval: Report and Results. In *Proceedings of the Second International Conference on Language Resources and Evaluation*: 1239-1244.

Lesk M. (1986). Automatic sense disambiguation: How to tell a pine cone from an ice cream cone. In *Proceeding of the 1986 SIGDOC Conference*. Association for Computing Machinery: 24-26.

Manning C.D. and Schütze H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Ng H. and Lee H. (1996). Integrating mutliple knowledge sources to disambiguate word sense: an exemplar-based approach. In *Proceedings of 34th Annual Meeting of the Society for Computational Linguistics*: 40-47.

Schütze H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, vol. (24): 97-124.

Yarowsky D. (1992). Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics*: 189-196.

Yarowsky D. (1994). Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*: 88-95.

Defeating the Homogeneity Assumption

Anne De Roeck, Avik Sarkar, Paul H. Garthwaite

Faculty of Maths and Computing – The Open University – Walton Hall – Milton Keynes – MK7 6AA, U.K.
{a.sarkar; a.deroeck; p.h.garthwaite}@open.ac.uk

Abstract

The statistical NLP and IR literatures tend to make a "homogeneity assumption" about the distribution of terms, either by adopting a "bag of words" model, or in their treatment of function words. In this paper we develop a notion of homogeneity detection to a level of statistical significance, and conduct a series of experiments on different datasets, to show that the homogeneity assumption does not generally hold. We show that it also does not hold for function words. Importantly, datasets and document collections are found not to be neutral with respect to the property of homogeneity, even for function words. The homogeneity assumption is defeated substantially even for collections known to contain similar documents, and more drastically for diverse collections. We conclude that it is statistically unreasonable to assume that word distribution within a corpus is homogeneous. Because homogeneity findings differ substantially between different collections, we argue for the use of homogeneity measures as a means of profiling datasets.

Keywords: homogeneity, term distribution, corpus profiling.

1. Introduction

It is common practise in some areas of statistical Natural Language Processing (NLP), and Information Retrieval (IR), to assume that terms in a document occur independently of each other. This gives rise to the well known "bag of words" model for text which, in spite of numerous drawbacks (see Franz, 1997), has been used extensively. One of the major reasons behind this is the success of the vector-space approach to IR. The model also makes the application of standard mathematical and statistical techniques very convenient. At the same time, it is widely accepted that the term independence assumption is wrong, and that words do not occur independently of each other. Though some information retrieval techniques are said to work precisely because language "mildly" defeats the assumption (Manning and Schütze, 2000), the actual extent to which the occurrence of terms depends on other terms is relatively unexplored.

Term independence is related to the notion of homogeneity in term distribution. When text is seen as a "bag of words", terms are expected to distribute evenly throughout documents. Yet they do not. There is a growing literature which investigates "burstiness" in the distribution of content words in documents - i.e. the fact that repeated occurrences of an informative word in a document tend to cluster together (eg Church, 1995). By and large, however, function words are ignored, or assumed to distribute evenly throughout text, to the point of becoming uninformative. Indeed, Katz (1995) develops a model for bursty distributions of "concept terms", and distinguishes between function words and content words on the grounds that function words are distributed more homogeneously throughout text.

In short, the statistical NLP and IR literatures tends to make a "homogeneity assumption", either as a consequence of a "bag of words" model, or in the treatment of function words. In this paper, we show that the homogeneity assumption does not generally hold. In particular, we show that it does not hold for function words. Secondly, we show that datasets and document collections display different characteristics with respect to (non-)homogeneity, even when based on function words. Specifically, we show that the homogeneity assumption is defeated substantially for collections known to contain similar documents, and even more drastically for diverse collections. Evidence of homogeneity in term distribution rarely survives beyond very small text chunks.

We start from work in the corpus literature, which casts homogeneity as a property of term frequency distributions. Using Kilgariff's (1997) methodology as a base line, we use the χ^2 test (including the p-value) to relate a notion of homogeneity to a level of statistical significance. We explore different ways of partitioning datasets and investigate homogeneity in a range of collections with different characteristics. We also design, and report on, experiments where we investigate the effects of inducing different levels of randomness in text drawn from different collections.

Our results are significant in their own right, in that they show that it is statistically unreasonable to assume that word distribution within a corpus is homogeneous. In addition, in showing that document collections are not neutral with respect to the property of homogeneity, we make an indirect argument for using homogeneity measures to profile textual datasets. Such measures would help application developers to estimate the differences (including genre differences) between datasets. They would also help evaluation exercises, where experimental results can be supplemented with information on the characteristics of the dataset on which they were obtained.

2. Homogeneity

Homogeneity in document collections has been approached from different perspectives. We will concentrate on homogeneity as a property of term frequency distribution, or word count. Kilgariff (1997) describes a basic method for using measures of similarity to gauge homogeneity in a corpus. Starting from the position that no corpus can be more similar to another corpus than it is to itself, he casts homogeneity as internal similarity of distributions, between two halves of a document collection. Clearly, distributions of different features can be checked for similarity. Primarily interested in language variety, he proposes to measure term frequency distributions, and initially uses the χ^2 statistic. In the corpus literature, measuring this particular flavour of homogeneity has been linked to gauging the distance between corpora and genre detection. Rose and Haddock (1997) suggest using similarity based homogeneity measures to verify language model quality in corpus acquisition. Rayson and Garside (2000) show applications in the study of social differentiation in the use of English. Cavaglia (2002a) defines a homogeneous corpus as one that belongs to the same sub-language. In all these, focus tends to be on similarity as a means of establishing that two collections belong to the same genre or sub-language, by measuring lexical and syntactic features such as term frequency or POS distributions. A departure from this theme is Cavaglia (2002b), who uses term frequency and POS distributions together with unsupervised learning to generate corpora. Cavaglia (2001) uses homogeneity measures on web documents to judge the spread of documents based on certain keyword searches.

2.1. *How to measure homogeneity in term frequency distribution?*

Kilgariff (1997) sees homogeneity as internal similarity. His basic method for measuring homogeneity involves five steps:

(1) Divide the corpus into two halves by randomly placing text in one of two sub-corpora;

(2) Produce a word frequency list for each sub-corpus;

(3) Calculate the χ^2 statistic for the difference in term frequency distributions between the two sub-corpora;

(4) Normalise for corpus length;

(5) Iterate over successive random halves.

Kilgariff (1997) and Rose and Haddock (1997) partition their corpus by placing successive chunks of 5000 words in each half. The basic technique of comparing two halves of a corpus has been used with different similarity measures. Kilgariff (1997) adopts χ^2, Rose and Haddock (1997) and Rayson and Garside (2000) use G2. Alternatives include correlation on term rank frequency data, such as Mann-Whitney (Kilgariff, 1996) or Spearman's S (Rose and Haddock, 1997). Kilgariff and Rose (1998) compare Spearman's S with χ^2. Rayson and Garside (2000) deploy log-likelihood on different features, to expose different aspects of similarity. Cavaglia (2002b) uses relative entropy (Kullback-Leibler divergence), χ^2 and G2.

χ^2 is found to perform well in comparative experiments (Cavaglia, 2002b; Rose and Haddock 1997), as long as certain conditions are met. Notably, each of the individual frequency values must be greater than or equal to 5. Dunning (1993) states that most statistical tests assume some underlying distribution (usually either normal or Chi-Square (χ^2)). He shows through experiments that these assumptions can only be made if the sample size is large enough. He also discusses likelihood ratio tests and compares the results with that of Pearson's χ^2 test.

2.2. *The χ^2 Test and the χ^2 statistic*

At this point, it is important to clarify the relationship between the χ^2 test and the χ^2 statistic. The χ^2 test is a standard method to test the hypothesis that two or more samples are homogeneous, i.e. that they are drawn from the same population at random. In the SPlus software on a UNIX platform which we used for the experiments, the χ^2 test has three values in the output. First, the χ^2 statistic is calculated by the following formula:

$$\chi^2 = \Sigma ((O-E)^2/E)$$

where it tests the difference between expected (E) and observed (O) occurrences of events and is calculated with (N-1) degrees of freedom (this is the second output). N is the number of terms under consideration. The third output value is the p-value, a measure of confidence as to whether the two samples are statistically significant. The p-value is actually a probability depicting the level of confidence about the judgement based on the sample size. Being a probability, its value lies in the range 0 to 1. A value close to 0 indicates that, based on the sample size, the null hypothesis of similarity between two samples should be rejected.

The χ^2 statistic, on the other hand, has also been seen as a similarity measurement. In the case of perfect similarity (i.e. homogeneity in our case), one would expect the observed and expected occurrences to be close. Hence a lower χ^2 value would indicate greater similarity as compared to a higher χ^2 value. As a consequence, the χ^2 value may be viewed as a measure for comparing the similarity of two corpora, provided the degrees of freedom (N-1) is kept constant. This is due to the fact that a χ^2 value is calculated by summation over all the terms

under consideration, which leads to a higher value if more terms are considered. The effect of number of terms considered can be approximately nullified by dividing the χ^2 value by the degrees of freedom (N-1). The measure is called Chi-square By Degrees of Freedom (or *CBDF*). This is the corpus homogeneity measure used by Kilgariff (1997). Most other work (Kilgariff, 1996 and 1997; Rayson and Garside, 2000; Rose and Haddock, 1997) on corpus homogeneity also deploys the χ^2 statistic as a measure, rather than as a statistical test of significance.

Even a small departure from homogeneity can be detected if a sample's size is large enough, the p-value will get closer and closer to 0 as the sample size increases. One would like a measure of homogeneity that is not affected greatly by sample size, so that corpora of different lengths can be compared. Also, it is preferable if the similarity measure is compatible with a test of homogeneity: if two corpora are of similar size, the one with the larger value on the similarity scale should also have the smaller p-value for the test of homogeneity. Using CBDF as the similarity measure and the χ^2 test as the test of homogeneity gives these desirable properties.

3. Experimental Framework

3.1. *Homogeneity detection to a level of statistical significance*

Our aim of investigating the homogeneity assumption requires a more fine-grained tool than simple use of the χ^2 statistic as a homogeneity measure. We are interested in conditions under which non-homogeneity is detected, and in factors that affect the degree of non-homogeneity in different datasets.

We will adopt Kilgariff's outline methodology described in section 3.1, and conduct our experiments based on χ^2, because it is found to perform well in comparative experiments (Cavaglia 2002b; Rose and Haddock 1997), as long as certain conditions are met. (In particular, each of the individual frequency values must be greater than or equal to 5.) The settings in which we have conducted our experiments satisfy these criteria.

We will differentiate results in two ways by reporting the p-value as well as the CBDF statistic. Given a null hypothesis (in our case, homogeneity), the p-value allows us to estimate the strength of the evidence offered by the data. A p-value < 0.1 is usually interpreted as constituting weak evidence against the hypothesis, a p-value < 0.01 as strong evidence against, and $p< 0.001$ as very strong evidence against the hypothesis. Normally, a p-value < 0.05 is considered significant (moderate evidence against the hypothesis). In our case, a p-value < 0.05 will be taken to indicate that non-homogeneity is statistically significant. The CBDF measure relates to the text and indicates the level of heterogeneity.

3.2. *Now divide a corpus*

Kilgariff's basic method (section 3.1) requires a corpus to be split into two halves, by randomly placing text in one of two sub-corpora. The obvious question is how to execute this division? One way might be to dissolve document boundaries and split the corpus halfway. Kilgariff (1997) and Rose and Haddock (1997) dissolve document boundaries, but place consecutive chunks of 5000 words in each partition. Why chunks of size 5000 were chosen, rather than some other size, is not explained. The method of partitioning a document set raises important questions that may affect the outcome of similarity based experiments. A chunk size of 1, for example, would give randomness, which we would expect to see reflected in the experimental results. Also, can chunk size be chosen independently of the document sizes, or

genres, in a corpus? What are the implications for homogeneity experiments if chunks of varying sizes are considered?

To answer some of these questions, we experimented with alternative ways of partitioning a corpus, with different ways of handling document boundaries. We also investigated a range of smaller chunk sizes. Briefly, we conducted three experiments:

1. Choose a document and assign it at random to either of two partitions (docDiv experiment).
2. Divide each document in the middle, and randomly assign one half to either of the partitions, and the other half to the other partition (halfdocDiv experiment).
3. Remove document boundaries and repeat the same experiments of Kilgariff (1997) with various chunk sizes, from 5 to 5000, and observe the homogeneity measure (chunkDiv experiment).

Kilgariff (1997) measures homogeneity using all terms which occur more than 5 times in each of the partitions. Since the homogeneity measures we are deploying are based on word count, the inclusion of the most frequent terms means that the behaviour of function words will dominate the outcome of our experiments, and our measure of homogeneity is examining largely stylistic homogeneity.

To allow more detailed tracking of the distribution of very frequent terms, we will, for each experiment, report results at different values for N. Experimental results are shown in Tables 4 to 7. CBDF and p-values are averaged over iterations.

4. Datasets

We aim to investigate homogeneity in datasets with different characteristics, and considered corpora of various types and stylistic differences.

Data Set	Contents of the documents
AP	Copyrighted AP Newswire stories from 1989.
DOE	Short abstracts from the Department of Energy.
FR	Issues of the Federal Register (1989), reporting source actions by government agencies.
PAT	U.S. Patent Documents for the years 1983-1991.
SJM	Copyrighted stories from the San Jose Mercury News (1991).
WSJ	Copyrighted stories from the Wall Street Journal (1987-1989).
ZF	Information from the Computer Select disks for 1989/1990, copyrighted by Ziff-Davis Publishing Co.
OU	The Open University intranet and extranet web-pages.

Table 1. Description of content of each of the datasets

We selected the seven different datasets of the TIPSTER collection. Apart from availability, and its use as an evaluation and benchmarking standard, this collection has other advantages for our purposes. Table 1 lists the datasets and shows they are artificially compiled, with some drawn from a narrow base of similar text types, or from a particular domain. To contrast our results, we also experimented on data collected from the Open University Intranet. This data-

set is more diverse in terms of document type and domain content than the TIPSTER ones. Table 2 gives some basic profiling statistics, which show some of the bias in the datasets. DOE, for example, appears relatively uniform regarding text length, whereas FR shows the largest range. Comparing the ratio of new to old words gives an indication of domain diversity. There is a significant difference between the rate of new terms occurring, between the OU dataset (1 in 131 words) and the SJM dataset (1 in 260 words), in spite of their similar size. The WSJ and SJM sets are quite close in size and characteristics as well as in genre type, so we would expect them to behave in similar ways. Note also that the 10 most frequent terms of all TIPSTER collections are function words, but not in the OU dataset (Table 3).

Data Set	No of Docs	Corpus Length (words)	Average Doc Length (words)	No of Distinct Terms	Average Distinct Terms per Doc	Shortest Doc (words)	Longest Doc.
AP	242,918	114,438,101	471.1	347,966	238.25	9	2,944
DOE	226,086	26,882,774	119.0	179,310	72.90	1	373
FR	45,820	62,805,175	1,370.70	157,313	292.65	2	387,476
PAT	6,711	32,151,785	4,790.91	146,943	653.05	73	74,964
SJM	90,257	39,546,073	438.15	178,571	223.60	21	10,393
WSJ	98,732	41,560,108	420.94	159,726	204.26	7	7,992
ZF	293,121	115,956,732	395.59	295,326	168.42	19	75,030
OU	53,681	39,807,404	744.36	304,468	219.87	1	15,430

Table 2. Basic profiling statistics of each of the datasets.

Data Set	10 Most Frequent Terms
AP	the, of, to, a, in, and, said, s, for, that.
DOE	the, of, and, in, a, to, is, for, with, are..
FR	the, of, to, and, a, in, for, or, that, be.
PAT	the, of, a, and, to, in, is, for, said, as.
SJM	the, a of, to, and, in, s, for, that, is.
WSJ	the, of, to, a, in, and, s, that, for, is.
ZF	the, m, p, and, to, of, a, in, is, for.
OU	the, of, to, a, and, j, in, k, is, report.

Table 3. 10 Most frequent terms in each dataset

5. Experimental results

5.1. docDiv

The docDiv experiment maintains document boundaries and compares similarity of the two halves after assigning whole documents randomly to either partition. This experiment investigates homogeneity across documents in a collection. As Table 4 (and Figure 1a) shows, the

experiment finds non-homogeneity ($p<0.05$) in almost all cases. The exceptions are the AP and the DOE datasets when the 10 and 20 most frequent terms are used, and the WSJ and SJM datasets for the 10 most frequent terms. All the other datasets show statistical significance, with p-values of 0 or close to it (very strong evidence against the homogeneity hypothesis). CBDF values provide further insight into the corpus. In most cases, they are quite large, indicating high levels of non-homogeneity.

5.2. halfdocDiv

The halfdocDiv experiment induces a level of randomness in the individual documents, by dividing each of the documents exactly halfway and assigning each half to one of the partitions. This experiment is sensitive to homogeneity within documents. Again, there was evidence of non-homogeneity between the two partitions.

Dataset	**Number of Terms (N)**							
	10	**20**	**50**	**100**	**500**	**1000**	**7000**	**20000**
AP	**2.107** **0.1216**	**1.576** **0.2139**	2.583 0.0003	2.290 0	2.732 0	2.601 0	2.441 0	2.435 0
DOE	**1.172** **0.463**	**1.450** **0.160**	1.755 0.0259	1.983 0	1.838 0	1.786 0	1.795 0	1.872 0
FR	54.524 0	41.715 0	72.093 0	66.787 0	51.387 0	61.266 0	39.043 0	23.534 0
PAT	21.074 0	29.315 0	62.494 0	55.353 0	50.265 0	44.824 0	32.056 0	22.468 0
SJM	**3.595** **0.1193**	2.768 0.0077	3.231 0	2.976 0	3.012 0	2.959 0	2.560 0	2.511 0
WSJ	**2.358** **0.178**	2.663 0.0019	2.364 0	2.335 0	2.623 0	2.749 0	2.831 0	2.917 0
ZF	11.947 0	8.133 0	6.907 0	6.576 0	6.122 0	5.634 0	4.595 0	4.576 0
OU	232.913 0	158.520 0	94.749 0	67.293 0	32.663 0	25.181 0	14.224 0	8.297 0

Table 4. docDiv Results. Average CBDF and p-value for a dataset using the N most frequent terms. Values in bold indicate cases where the homogeneity assumption has not been defeated (p>0.05).

Dataset	Number of Terms (N)							
	10	**20**	**50**	**100**	**500**	**1000**	**7000**	**20000**
AP	**1.774** **0.087**	**1.473** **0.117**	**1.369** **0.057**	**1.271** **0.066**	1.171 0.021	1.187 0.0001	1.147 0	1.136 0
DOE	**0.728** **0.655**	**0.931** **0.533**	**1.054** **0.438**	**1.043** **0.372**	**1.061** **0.195**	**1.027** **0.285**	**1.014** **0.271**	**1.01** **0.182**
FR	7.905 0.001	9.549 0	11.627 0	11.642 0	8.847 0	8.166 0	6.543 0	5.336 0
PAT	20.360 0	15.568 0	16.017 0	11.886 0	7.694 0	6.243 0	5.102 0	4.611 0
SJM	**1.323** **0.3860**	**1.569** **0.3919**	**1.320** **0.4436**	**1.469** **0.1069**	1.332 0	1.297 0	1.240 0	1.242 0
WSJ	**1.563** **0.279**	**1.618** **0.248**	**1.342** **0.203**	**1.298** **0.260**	1.236 0.017	1.210 0.0007	1.178 0	1.150 0
ZF	**1.948** **0.1288**	**1.858** **0.116**	1.709 0.0283	1.609 0.0240	1.559 0	1.598 0	1.536 0	1.556 0
OU	7.721 0.033	6.103 0.0025	8.091 0	8.216 0	6.366 0	5.502 0	4.223 0	3.087 0

Table 5. halfdocDiv results. Average CBDF and p-values for a dataset using the N most frequent terms. Values in bold indicate cases where the homogeneity assumption has not been defeated (p>0.05).

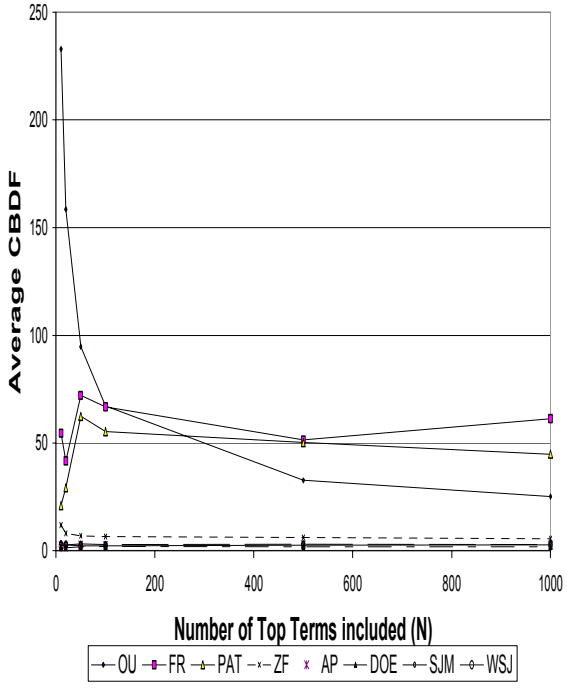


Figure 1a. docDiv

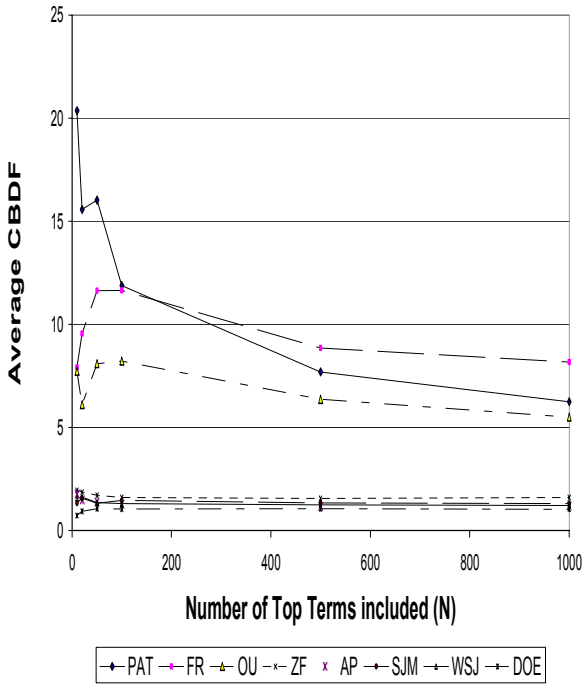


Figure 1b. halfdocDiv

Figure 1. (a) docDiv and (b) halfdocDiv for each dataset: Relationship between CBDF values and the N most frequent terms on which it is based.

Dataset	Number of Terms (N)							
	10	20	50	100	500	1000	7000	20000
AP	**0.628**	**0.836**	**0.871**	**0.984**	**0.990**	**1.007**	**1.018**	**1.012**
	0.7516	**0.6375**	**0.677**	**0.484**	**0.535**	**0.523**	**0.1595**	**0.179**
DOE	**1.141**	**1.225**	**1.151**	**1.050**	**1.038**	**1.002**	**1.008**	**1.008**
	0.3946	**0.3461**	**0.2505**	**0.3540**	**0.4229**	**0.462**	**0.431**	**0.3667**
FR	**0.754**	**0.961**	**0.967**	**1.033**	**1.016**	**1.025**	**1.022**	**1.013**
	0.650	**0.504**	**0.54**	**0.405**	**0.4174**	**0.335**	**0.2281**	**0.211**
PAT	**1.284**	**1.457**	**1.255**	**1.153**	**1.051**	**1.007**	**1.008**	**1.020**
	0.2451	**0.091**	**0.2273**	**0.1862**	**0.226**	**0.429**	**0.330**	**0.077**
SJM	**1.204**	**1.175**	**1.226**	**1.127**	**0.979**	**1.004**	**1.012**	**1.010**
	0.429	**0.375**	**0.293**	**0.268**	**0.608**	**0.454**	**0.262**	**0.181**
WSJ	**0.834**	**1.008**	**0.778**	**0.924**	**0.957**	**0.984**	**1.000**	**1.01**
	0.573	**0.492**	**0.822**	**0.679**	**0.682**	**0.6202**	**0.498**	**0.252**
ZF	**0.861**	**0.791**	**0.939**	**0.913**	**0.994**	**1.012**	**1.007**	**1.016**
	0.5781	**0.704**	**0.636**	**0.703**	**0.525**	**0.394**	**0.393**	**0.1258**
OU	**1.242**	**1.257**	**1.165**	**1.023**	**1.081**	**1.054**	1.042	1.033
	0.3395	**0.271**	**0.234**	**0.424**	**0.118**	**0.142**	0.034	0.005

Table 6. chunkDiv results with chunk size 5. Average CBDF values and p-values for a dataset using the N most frequent terms. Values in bold indicate cases where non-homogeneity is not statistically significant (p>0.05).

However, the experiment finds statistically insignificant non-homogeneity (p>0.05) much more often than the earlier docDiv experiment, with p-values higher than 0.05 for certain instances in more than half the datasets (Table 5; Figure 1b). Note that the DOE collection contains very short documents, each unlikely to deal with more than one topic. Also, CBDF values are much lower here than in the corresponding docDiv table. This suggests that terms distribute more homogeneously within documents, than across document boundaries. At the same time, with the 10 most frequent terms, there was evidence of heterogeneity among half-document partitions for three of the eight corpora, showing that, in general, even very frequent terms cannot be assumed to be uniformly distributed within a document. Hence this measure of homogeneity may be used to detect term burstiness in documents.

5.3. chunkDiv

Imagine dividing a dataset by randomly assigning each consecutive word to one of two partitions. Such a division would give randomness in the partitioning, and destroy any evidence of dependencies between terms. In this case, we would expect our experiments to seldom register statistically relevant evidence of non-homogeneity (we confirmed this experimentally for these datasets). On the other hand, Kilgariff (1997) reports non-homogeneity in partitions assigning chunks of 5000 words, which give far less randomness. Two questions arise. For a particular dataset, how large must the chunks be before non-homogeneity in the distribution of terms is statistically significant (p<0.05)? Is this level dataset dependent?

The chunkDiv experiment is designed to investigate the effects of different levels of randomness in a partitioning. We merged each dataset into a single document as in Kilgariff (1997), but placed a series of smaller chunks in partitions, ranging from 1 to 5000. We report only on chunk sizes 5 (Table 6; Figure 2a) and 100 (Table 7; Figure 2b).

Dataset	Number of Terms (N)							
	10	**20**	**50**	**100**	**500**	**1000**	**7000**	**20000**
AP	**0.824**	**1.105**	**1.412**	1.607	1.471	1.372	1.3004	1.3026
	0.6023	**0.3560**	**0.0735**	0.0019	0	0	0	0
DOE	**1.102**	1.864	1.646	1.511	1.354	1.414	1.4013	1.424
	0.3937	0.0280	0.0231	0.0317	0.0299	0	0	0
FR	**1.006**	**1.441**	**1.608**	1.803	1.924	1.834	1.782	1.746
	0.5071	**0.229**	**0.076**	0.025	0	0	0	0
PAT	4.181	3.051	2.682	2.420	2.252	2.104	1.977	1.876
	0.0232	0.0025	0.0007	0	0	0	0	0
SJM	**0.995**	**1.117**	**1.146**	**1.180**	1.410	1.402	1.317	1.291
	0.4720	**0.3851**	**0.3203**	**0.2463**	0	0	0	0
WSJ	**1.112**	**1.213**	**1.198**	**1.230**	1.196	1.283	1.2902	1.319
	0.3741	**0.324**	**0.2426**	**0.0937**	0.0383	0	0	0
ZF	**1.576**	**1.283**	1.709	2.190	1.41	1.673	1.315	1.884
	0.4152	**0.366**	0.011	0	0	0	0	0
OU	6.231	5.657	4.870	4.278	3.310	2.733	2.261	1.865
	0.0004	0	0	0	0	0	0	0

Table 7. chunkDiv results with chunk size 100. Average CBDF values and p-values for a dataset using the N most frequent terms. Values in bold indicate cases where non-homogeneity is not statistically significant (p-value>0.05).

Our results show a systematic relationship between increasing chunk size and increasing non-homogeneity: in Table 7, there are fewer non-significant p-values than in Table 6, and the CBDF values are higher. There also appears to be a relationship between registering non-homogeneity and a combination of document length and diversity of domain coverage. Where a dataset contains many very short documents, even small chunks are likely to cross document boundaries. Where such collections also cover diverse domains, documents are more likely to contain a higher proportion of distinct terms for the same amount of text. This would explain why the OU data started registering non-homogeneity at smaller chunk sizes than the other collections, as it combines a high incidence of short documents with diverse domain coverage.

Where they are function words, high frequency terms require bigger chunk sizes before non-homogeneity is apparent, when compared to experiments with more (less frequent) terms. Also CBDF values are lower when only high frequency terms are considered. To some extent, these results confirm Kilgariff (1996) and Katz (1995) who anticipate that more frequent function words have more similar distributions among documents than less frequent (content)

terms. Importantly, however, there are clear differences between the behaviour of function words in different datasets of the TIPSTER collection. (Results for the OU dataset are consistent with the conjecture of Kilgariff and Katz, because the OU most frequent terms contain non-function words). In all, the chunkDiv experiment revealed that the distribution of function words is very different from the distribution of content words.

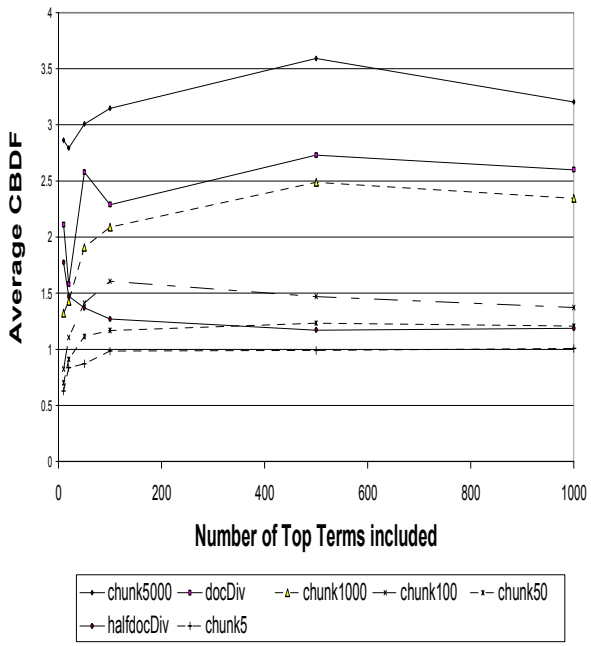


Figure 2a. AP Dataset

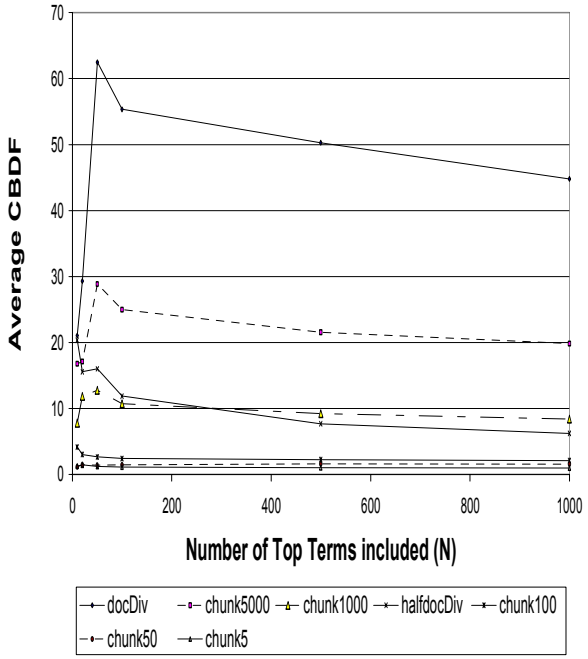


Figure 2b. PAT Dataset

Figure 2. (a) AP Dataset and (b) PAT Dataset. Relationship between CBDF values and N most frequent terms for all partitions (docDiv, halfdocDiv, and various chunk sizes).

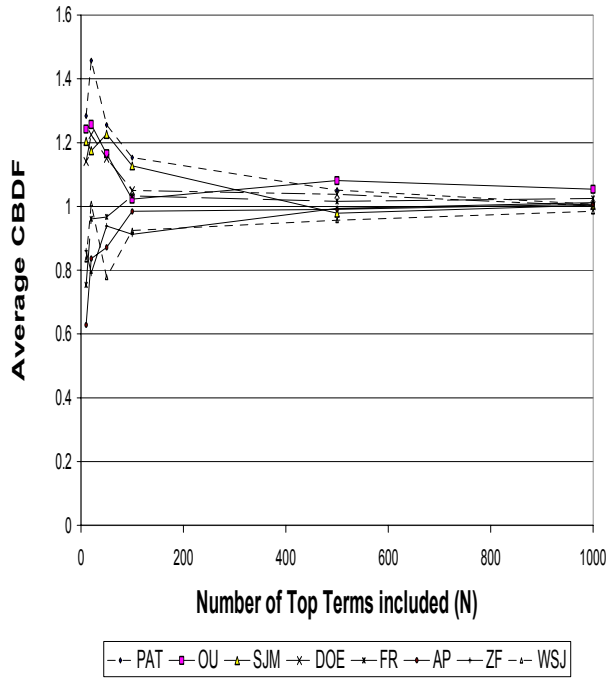


Figure 3a. chunksize 5

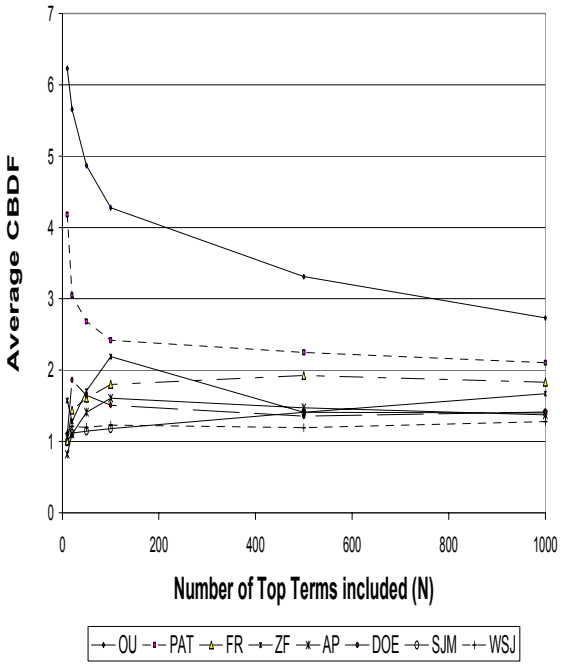


Figure 3b. chunksize 100

Figure 3 chunkDiv for (a) chunksize 5 and (b) chunksize 100 for each dataset: Relationship between CBDF values and the N most frequent terms on which it is based.

We plot average CBDF values of the docDiv. halfdocDiv and the various chunkDiv values for two datasets only, PAT and AP (Figure 2), due to lack of space. We show CBDF values using N most frequent terms up to N=1000. (There is not much variation for higher values of N.)

The figures show significant differences between the two datasets; the PAT dataset is much less homogeneous than the AP dataset. One may also note the increase in average CBDF value with increasing chunk sizes, which indicates increased non-homogeneity.

To show that collections have different homogeneity properties, we plot the results for all the datasets simultaneous for chunks of size 5 (Figure 3a) and size 100 (Figure 3b).

6. Contribution and Limitations

We have refined the Kilgariff (1997) method for measuring homogeneity by introducing a more fine grained approach for estimating homogeneity properties of datasets. We have supplemented the use of the CBDF measure with an indication of the strength of the evidence in the data. We have also experimented with dataset partitioning to investigate at which point non-homogeneity becomes detectable, and used chunk sizes together with the p-value to gain a view on how quickly the homogeneity assumption is defeated. Our approach has given us an opportunity to look closely at the behaviour of very frequent words and function words, which varies considerable across datasets. Our ultimate goal is to establish a collection of measures whose values might inform the deployment of techniques appropriate to the profile of the data. Whilst homogeneity measures can be used to estimate distance between datasets of different genres, much further work is needed to identify how the properties of datasets affect measures and values. Also, in the absence of results pertaining to less frequent (non-function) words, the work is of limited benefit for applications which make use of stop lists.

7. Conclusions and Future Work

We have investigated homogeneity in term distribution of the N most frequent terms. Starting from Kilgariff's work, we developed a notion of detecting non-homogeneity with a level of statistical significance, and have experimented with different partitions of a range of datasets. We conclude that the homogeneity hypothesis does not generally hold, even for function words. We also showed that different datasets will exhibit different homogeneity properties, which appear to correlate with a range of characteristics of the dataset. We conclude that it is statistically unreasonable to assume without question that word distribution within a corpus is homogeneous. In analysis of a corpus it is often very convenient to treat term distribution as homogeneous, and whether results would be biased to an important extent will depend on the analysis being performed and the purpose for which it is required. Our results show that it will also depend on the corpus being analysed, because the degree of non-homogeneity differs substantially between different collections. We argue for the use of homogeneity measures as a means of profiling datasets, in part to decide if an assumption of homogeneity is likely to lead to serious error.

Our objectives for future work are to find models of word distribution that fit reality better than the homogeneity assumption, for very frequent terms (including function words), and integrate them with "burstiness" phenomena for less frequent terms. Our chunkDiv experiments showed that the most frequent terms introduce a high degree of variability in homoge-

neity results, and we have started to investigate similar experiments where very frequent terms have been disregarded. We also want to investigate further the extent to which homogeneity measures are useful practical tools for profiling datasets.

References

Cavaglia G. and Kilgariff A. (2001). Corpora from the Web. In *Proceedings of the 4th Annual CLUK Colloquium*, Sheffield, UK, January 2001.

Cavaglia G. (2002a). Measuring the homogeneity of different varieties of language. In *Proceeding of the 5th National Colloquium for Computational Linguistics in the UK (CLUK)*, Leeds: 37-44.

Cavaglia G. (2002b). Measuring corpus homogeneity using a range of measures for inter-document distance. ITRI Report Series, ITRI-02-08, University of Brighton, UK.

Dunning T. (1993). Accurate Methods for the Statistics of Surprise and Co-incidence. *Computational Linguistics*, vol. (19/1): 61-74.

Franz A. (1997). Independence Assumptions considered harmful. In *Proceedings of ACL 1997*: 182-189.

Church K. (2000). Empirical Estimates of Adaptation: The chance of Two Noriega's is closer to p/2 than p^2. In *Proceedings of Coling*: 173-179.

Katz S. (1996). Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, vol. (2/1): 15-59.

Kilgariff A. (1996). Which words are particularly characteristic of a text? A survey of statistical approaches. In *Proceedings AISB Workshop on Language Engineering for Document Analysis and Recognition*: 33-40.

Kilgariff A. (1997). Using word frequency lists to measure corpus homogeneity and similarity between corpora. In *Proceedings of ACL-SIGDAT Workshop on very large corpora*, Hong Kong.

Manning Chr. and Schuetze H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Rayson P. and Garside R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora*, *ACL 38*, Hong Kong: 1-6.

Rose T. and Haddock N. (1997). The effects of corpus size and homogeneity on language model quality. In *Proceedings of the ACL-SIGDAT workshop on very large corpora*, Hong Kong: 178-191.

Que faire des corpus multilingues parallèles ? Une expérience

Jean-Claude Deroubaix

Groupe de recherche sur les acteurs internationaux et leurs discours
(GRAID-Institut de Sociologie) – Université Libre de Bruxelles
44 av. Jeanne – 1050 Bruxelles – Belgique
deroubaix@swing.be

Abstract

With the movement going to a globalisation of political systems, we are confronted more and more with political text written in more than one language. This proliferation was first a challenge for translators. Some progress have been done to facilitate the manipulation of these texts in the context of translation. We propose to examine how it is possible to use traditional tools of lexicometrics to describe and analyse this kind of text from the point of view of the social scientist.. We'll examine also the benefits we could obtain from these attitude.

Résumé

Face au mouvement de mondialisation des systèmes politiques, nous sommes confrontés de plus en plus souvent à des textes politiques rédigés en plusieurs langues. Cette prolifération fut d'abord un défi pour les traducteurs. Des progrès ont été réalisés pour faciliter le traitement de ces textes dans le contexte de la traduction. Nous nous proposons d'examiner comment les outils classiques de la lexicométrie peuvent être utilisés pour décrire et analyser ce genre de corpus du point de vue des sciences sociales. Nous examinerons aussi les bénéfices que l'on peut attendre d'un tel point de vue.

Mots-clés : discours politique, statistique lexicale, corpus multilingues.

1. Les corpus multilingues

Lors des JADT[1] de Nice, ma communication (Deroubaix, 1998) avait porté sur la nécessité de construire des outils d'analyse performants des corpus multilingues. En effet, l'émergence puis le renforcement d'accords institutionnels et politiques qui dépassent largement le cadre de simples accords classiques de relations internationales diplomatiques pour constituer des sociétés politiques couvrant de grandes régions du monde (Union européenne, ALENA) ou plus radicalement le monde entier (OMC) engendrent dès lors des textes politiques qui s'imposent souvent comme supérieurs aux normes nationales. Ces textes sont généralement multilingues, ayant valeur légale, dans chacune des langues originelles (onze langues pour l'Union européenne par exemple et pour l'instant). En outre ils sont traduits dans d'autres langues si nécessaire en vue d'être compris par l'ensemble des populations qui y sont soumises.

La multiplication de tels textes soulève de nombreuses questions.

Les premières concernent essentiellement les traducteurs qui ont à faire face à une croissance exponentielle de texte et à la nécessité d'automatiser au moins partiellement leur travail,

[1] Pour la réalisation de cette communication nous avons utilisé la suite logicielle LEXICO1 pour Mac développée au Laboratoire de Lexicologie politique de Saint-Cloud par André Salem. Les analyses factorielles ont été réalisées avec les programmes de l'ADDAD, mis à la disposition des utilisateurs de LEXICO1.

d'une part, et d'autre part de maintenir une cohérence traductive difficile à obtenir lorsqu'il s'agit d'effectuer des traductions entre systèmes politiques parfois bien différents[2].

Une seconde liste de questions concernent plus particulièrement l'analyse de ces textes politiques auxquels il peut paraître un peu trop simple d'appliquer des analyses scientifiques sur une seule des versions linguistiques sans prendre en compte le fait qu'il ne s'agit en l'occurrence que d'analyse portant sur un extrait d'un corpus multilingue par nature.

Les questions proprement traductives ont suscité évidemment de nombreux travaux tant sur la manière de constituer des mémoires traductives que sur la réalisation semi-automatique de glossaires ou même pour aller en amont de ces deux démarches pour réaliser des outils d'alignement de corpus. (cf., par exemple, la bibliographie dans Kraif, 2001).

2. Les questions d'analyse lexicale

Un moindre intérêt a été porté à l'analyse lexicale multilingue. Pourtant, les outils classiques de l'analyse lexicométrique peuvent fournir des résultats intéressants lorsqu'ils sont appliqués à des corpus multilingues. En effet, peu de chose distinguent formellement les corpus multilingues des corpus traditionnellement explorés par les analyses lexicométriques. L'élément fondamental sur lequel les calculs statistiques de la lexicométrie vont être réalisés est le tableau lexical entier c'est-à-dire un tableau des fréquences des formes lexicales (ou des lemmes, éventuellement) attestées dans ce corpus et ventilées dans l'ensemble des parties du corpus.

C'est sur ce tableau que vont se calculer les spécificités, vont se construire des classifications des parties ou des formes lexicales ou se calculer divers indices de richesse de vocabulaire par exemple. Voici le début d'un tel tableau. Il s'agit de la distribution des occurrences des mots de plus grande fréquence dans le corpus constitué des 11 textes intitulés « Grandes orientations de politique économique »[3] adoptés par le Conseil européen. Il s'agit du corpus français.

GOPE	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
de	143	143	232	179	285	500	970	1381	2010	1927	1966
la	104	123	168	134	187	324	555	876	1232	1093	1256
des	84	78	150	111	163	282	523	668	979	852	966

Cependant, dans le cas d'un corpus multilingue, nous disposons d'autant de tableaux lexicaux qu'il y a de langues utilisées dans le corpus parallèle. Ainsi pouvons-nous construire :

En italien :

GOPE	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
di	105	101	143	111	218	402	726	1018	1331	1415	1384
e	69	75	98	95	139	203	413	658	839	977	897
la	52	69	83	62	96	132	257	442	632	669	656

[2] Ainsi est-il malaisé de traduire les concepts qui organisent la sécurité sociale de chacun des pays membres de l'UE sans agir sur leur signification. Cf. le sens de « cotisation sociale » dans Friot (2003).

[3] Ces documents portent évidemment des noms différents dans les versions linguistiques différentes. Nous désignerons désormais ce corpus, et les textes qui le constituent, avec l'abréviation française GOPE quelle que soit la langue considérée.

En anglais :

GOPE	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
the	202	201	280	239	368	527	981	1537	1913	2179	1934
of	90	79	145	113	173	254	484	818	1087	1226	1029
in	79	79	114	110	189	243	479	752	1074	1160	993

En allemand :

GOPE	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
die	138	146	155	135	195	322	579	711	1215	1282	1120
der	131	102	154	139	193	287	569	950	1044	1201	1175
und	71	77	110	110	150	236	505	672	934	1009	965

La juxtaposition de ces tableaux lexicaux nous donne un nouveau tableau lexical qui correspond à celui que nous aurions obtenu en soumettant à la segmentation et au comptage un corpus multilingue dont les parties seraient chacune constituée des différentes versions linguistiques du même texte disposées à la queue leu leu.

	GOPE	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
1	de	143	143	232	179	285	500	970	1381	2010	1927	1966
2	la	104	123	168	134	187	324	555	876	1232	1093	1256
3	des	84	78	150	111	163	282	523	668	979	852	966
…	…	…	…	…	…	…	…	…	…	…	…	…
1487	visent	0	0	1	0	0	0	1	1	2	2	3
1488	di	105	101	143	111	218	402	726	1018	1331	1415	1384
1489	e	69	75	98	95	139	203	413	658	839	977	897
1490	la	52	69	83	62	96	132	257	442	632	669	656
…	…	…	…	…	…	…	…	…	…	…	…	…
3066	vincolanti	0	0	0	0	1	0	3	1	3	2	0
3067	the	202	201	280	239	368	527	981	1537	1913	2179	1934
3068	of	90	79	145	113	173	254	484	818	1087	1226	1029
3069	in	79	79	114	110	189	243	479	752	1074	1160	993
…	…	…	…	…	…	…	…	…	…	…	…	…
4420	week	0	0	0	0	0	0	1	3	2	4	0
4421	die	138	146	155	135	195	322	579	711	1215	1282	1120
4422	der	131	102	154	139	193	287	569	950	1044	1201	1175
4423	und	71	77	110	110	150	236	505	672	934	1009	965
…	…	…	…	…	…	…	…	…	…	…	…	…
6017	zunehmende	0	0	0	0	0	2	0	1	2	4	1

En pratique, cependant, il importe de recourir à une segmentation séparée des textes selon la langue en vue d'éviter quelques problèmes d'homographies translinguistiques[4].

[4] Par exemple lors du traitement du corpus bilingue des déclarations gouvernementales belges (Deroubaix, 1997) nous nous sommes trouvé devant le problème, trivial, de devoir distinguer *de* préposition en français, du

L'intérêt d'une telle démarche réside dans la possibilité de donner comme point de référence aux analyses, la totalité du corpus multilingue et non celle d'une seule de ses parties. Ainsi dans l'application d'une analyse factorielle des correspondances, nous pouvons décrire le nuage des parties immédiatement par rapport à toutes les langues, nous pouvons aussi situer les différentes versions linguistiques de chacune des parties les unes par rapport aux autres. Quant au nuage des formes, il nous sera loisible entre autres de repérer les relations des formes entre elles indépendamment de leur langue. Ces avantages peuvent aussi être mis à profit lors de la construction de typologies par classification automatique.

3. Les GOPE et leurs formes les plus fréquentes

Dans nos démocraties politiques, il est de grands textes politiques qui suscitent les commentaires de la presse et font débat entre les citoyens ; parmi ces textes figurent certes les déclarations solennelles des gouvernements nationaux, les programmes de ces mêmes gouvernements et les programmes politiques des partis.. Les citoyens en prennent connaissance immédiatement ou par le biais des résumés fournis par la presse. Chacun d'entre-eux pressent que ces textes vont avoir une influence certaine sur la vie politique d'abord, sur la vie quotidienne ensuite. Les grandes orientations de politique économique adoptées par le conseil européen ne jouissent évidemment pas de la même publicité ni de la même attention du citoyen ou de la presse. Pourtant il s'agit de l'énoncé du programme de politique économique que l'UE attend de voir concrétiser dans l'année par les États membres. Ce programme est très concret : il contient des recommandations générales et des recommandations pays par pays. Les programmes gouvernementaux nationaux sont subordonnés au suivi de ces recommandations. C'est la raison pour laquelle, au GRAID, nous avons décidé d'étudier le vocabulaire de ces textes, car du fait même qu'ils se présentent sous la forme de recommandations, ils forment un pont, un vecteur de la circulation lexicale des termes du pouvoir (cf. Gobin, 2003), un lieu essentiel pour comprendre comment se dit et se fait la politique aujourd'hui.

Observons d'abord les 15 termes les plus fréquents de ces GOPE (en ayant éliminé de la liste les mots-outils, prépositions et articles essentiellement).

Formes lexicales les plus fréquentes dans les GOPE (avec leur rang)			
En français	*En italien*	*En anglais*	*En allemand*
22 emploi	19 lavoro	9 labour	27 öffentlichen
24 travail	25 mercato	13 market	29 maßnahmen
25 marché	26 crescita	14 growth	32 insbesondere
26 croissance	28 bilancio	17 employment	33 mitgliedstaaten
32 devrait	29 occupazione	19 economic	34 jahr
33 taux	35 stati	21 budgetary	36 bip
34 mesures	39 politiche	22 public	43 eu
35 budgétaire	40 membri	23 policy	44 2001
36 œuvre	41 particolare	27 government	49 2002
39 publiques	43 mercati	30 member	54 wirtschaft
40 marchés	44 tasso	32 states	56 2000

de article masculin ou féminin défini en néerlandais. En traitant séparément les deux corpus et en construisant le tableau lexical par juxtaposition, nous évitons ces confusions.

41 états	45 disoccupazione	33 markets	61 finanzen
42 politiques	46 pubbliche	34 gdp	64 jahren
43 membres	48 dovrebbe	36 measures	65 wachstum
46 chômage	50 livello	38 unemployment	67 unternehmen

Ce qui est frappant et ce quelle que soit la langue est l'importance centrale accordée à l'emploi et au travail dans ce qui n'est somme toute qu'un texte économique qui n'est pas sensé se pencher sur les politiques sociales. Des analystes ont déjà soulevé cette dépendance des politiques sociales européennes par rapport aux compétences économiques de l'Union, cependant la répétition brute des lexèmes travail et emploi dans ces GOPE confirme cette subordination. Les Grandes orientations de politique économique traite moins d'industries (=économie réelle) que de l'emploi comme variable d'ajustement des politiques économiques et monétaires.

Pourtant outre cette symétrie entre les différentes versions linguistiques, il convient de souligner aussi quelques divergences : le rang d'*emploi* et de *travail* ne sont pas identiques, le français se distinguant de l'anglais et de l'italien. Or nous savons, par l'étude des GOPE en français, que *travail* est fortement lié à *marché* par le syntagme « marché du travail » alors que la majorité des utilisations de *emploi* le trouve lié à *taux* dans le segment « taux d'emploi ». L'allemand du fait des déclinaisons qui multiplient les formes lexicales d'un même lemme et de son utilisation de mots composés soulève des difficultés plus grandes pour l'analyse des formes les plus fréquentes puisque un lemme comme *arbeit* n'apparaît sous l'espèce de *arbeitsmarkt* qu'au 113ème rang, *arbeitsmarkt* est rendu dans les autres langues par un polyforme comme *marché du travail*.

4. Une analyse des correspondances d'un corpus multilingue

Nous avons réalisé l'analyse des correspondances du tableau lexical du corpus multilingue (quatre langues) des GOPE en ne retenant que les formes dont la fréquence était égale ou supérieure à 10 c'est à dire 6017 formes.

GOPE	Nbre d'occurrences	Nbre de formes	Nbre de formes F>=10
français	160464	6442	1487
Italien	152665	7067	1579
anglais	133422	4746	1354
allemand	126108	10075	1597
corpus	572659	28330	6017

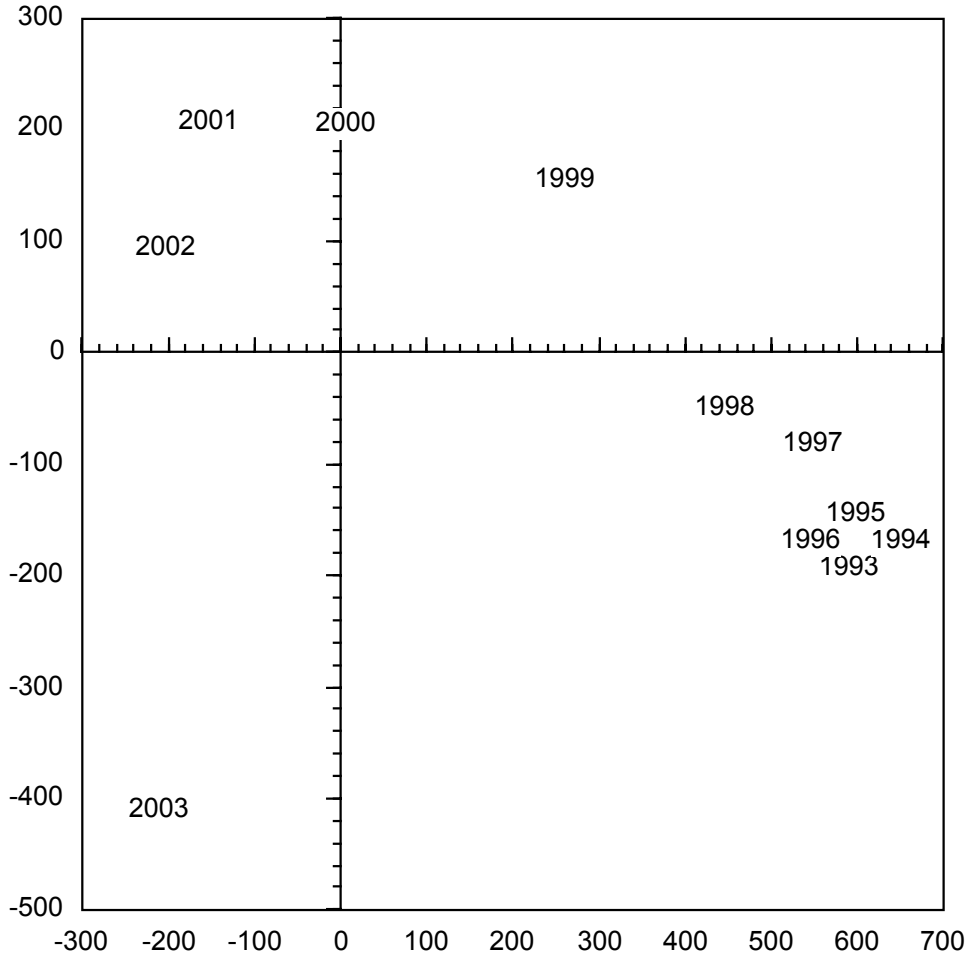
Le tableau lexical tronqué contient onze colonnes correspondant aux onze publications des GOPE de 1993 à 2003.

Le premier plan de l'AFC nous donne du nuage des GOPE une image en forme de parabole partant de 1993 en bas à gauche montant vers 2001 et redescendant ensuite dans le quadrant inférieur droit.

Nous savons (Deroubaix, 1997 ; Salem, 1997) qu'une telle image est la marque d'une série textuelle chronologique, c'est-à-dire d'un ensemble de discours dont le vocabulaire se renouvelle peu à peu, acquérant à chaque nouvelle publication une fraction de vocabulaire nouveau et en délaissant à chaque fois une autre fraction.

Il n'en reste pas moins qu'une lecture de ce plan factoriel du point de vue du vocabulaire ainsi acquis et rejeté est plein d'enseignement pour l'étude de la circulation lexicale et la modification des politiques énoncées. et qu'il est possible aussi d'examiner les ruptures et les écarts à la « parabole » dessinée sur le plan (Deroubaix et Gobin, 1987).

Ainsi, si l'ensemble des GOPE comprennent à la fois des recommandations générales (européennes) et des recommandations par pays, la manière dont sont réparties ces deux types de recommandations dans les textes et le poids qui leur est accordé ont varié dans le temps. Dans une première période de 1993 à 1998, l'organisation des recommandations est thématique (marché du travail, politique monétaire, …) et dans chaque thème sont reprises les recommandations générales et particulières.



À partir de 1999, le document comporte deux parties distinctes : l'une consacrée aux recommandations générales et l'autre aux recommandations détaillées pays par pays. Le poids des recommandations particulières est croissant. L'aspect normatif, et disciplinaire prend plus d'importance. Ce qui explique la rupture visible sur le plan factoriel entre le groupe 1999-1998 et celui de 1999 à 2002.

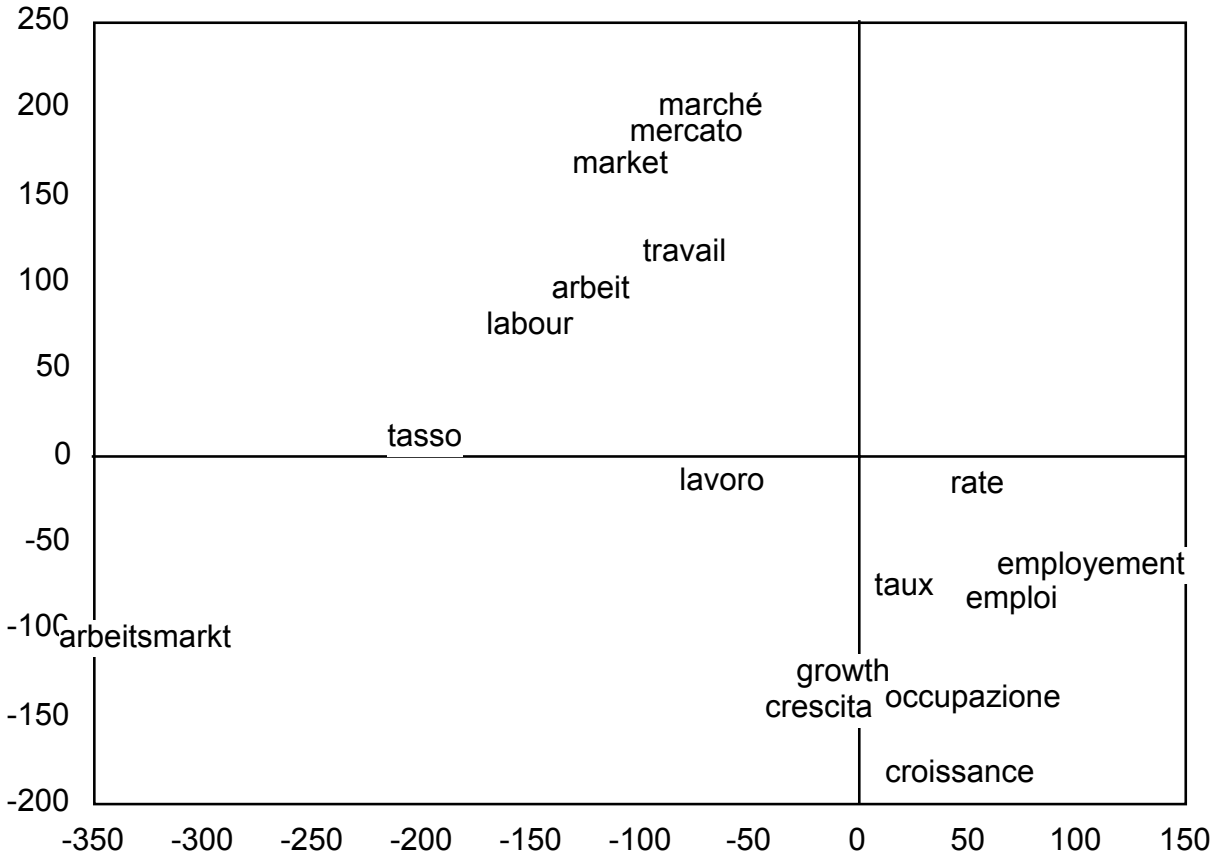
Avec cette AFC nous disposons d'un référentiel commun dans lequel il va être possible de projeter aussi les quatre versions linguistiques du corpus. Nous avons à cet effet ajouté aux 6017 lignes du tableau lexical, quatre lignes supplémentaires correspondant à la somme des occurrences des formes dans chaque langue (réduite de la même manière aux formes appa-

raissant au moins dix fois). La projection des profils « langue » nous montre que globalement les versions anglaise, française et italienne se démarquent peu du profil moyen.

Il n'en est pas de même du profil allemand, laissant à penser qu'au moins un découpage des formes composées devrait être effectué avant d'élaborer une comparaison.

Une exploration du lexique des GOPE en français (Gobin, 2003) a mis en évidence l'aspect central que ces documents donnent au « marché du travail », au « taux d'emploi » et à la « croissance », alors que l'on aurait pu s'attendre dans un texte intitulé « Grandes orientations de politique économique » à voir mis en valeur l'investissement, l'entreprise, etc. Nous nous sommes dès lors intéressé à la représentation des formes *marché*, *emploi*, *travail*, *croissance* et *taux,* et de leur traduction courante dans le corpus multilingue.

On constate tout d'abord que ces formes restent sensiblement groupées, d'une part le groupe *croissance*, *taux* et *emploi*, de l'autre le groupe *marché* et *travail*. On voit aussi qu'il y a eu un déplacement de l'accent mis sur ces deux groupes de formes entre la première période 93-98 et la seconde. Ceci ne signifie pas un désintérêt pour la croissance du taux d'emploi mais une montée relative de l'intérêt pour la réforme du « marché du travail ».



Cependant, cette homogénéité des représentations des formes dans les différentes langues est loin d'être absolue, *lavoro* mais plus encore *tasso* s'écartent du modèle moyen laissant supposer une utilisation différente.

5. Conclusions provisoires

Cette communication est une première approche, elle est partie d'un projet plus vaste d'étude de la circulation lexicale entre institutions internationales et nationales. Le corpus GOPE a été

choisi comme exemple de texte fondamental de politique économique et sociale de l'UE. En tant que tel il fait partie d'un ensemble de textes à partir desquels nous tenterons d'établir un glossaire de la protection sociale en Europe. Cela nécessite une exploration approfondie de ce corpus. La réalisation d'une typologie multilingue des formes et celle d'une décomposition de l'effet chronologique, selon une méthode mise au point dans (Deroubaix, 1997), sont en cours.

D'une certaine manière ces résultats encore préliminaires pourraient conforter le choix souvent opéré de travailler sur les versions unilingues de textes internationaux en vue d'explorer leur vocabulaire puisque les images des parties du corpus semblent relativement bien coïncider dans l'espace factoriel de référence. Pourtant les écarts observés entre les représentations des formes (pourtant parmi les plus fréquentes) selon la langue impose d'être prudent dans l'exportation pure et simple des conclusions tirées de l'étude d'une seule version linguistique. Les capacités informatiques actuelles et les outils statistiques permettant de traiter les corpus multilingues, il nous semble important de sauter le pas et travailler lorsque c'est possible sur les corpus les plus complets possibles. Nous appliquons ainsi d'ailleurs un des principes de l'étude sur corpus : l'exhaustivité.

Annexe

Pour la lisibilité des plans factoriels certains points ont été légèrement déplacés. Voici donc les coordonnées exactes des 11 GOPE.

J1	QLT	POID	INR	1#F	COR	CTR	2#F	CTR
1993	365	16	71	603	262	75	-178	2
1994	423	17	76	648	311	95	-190	1
1995	498	24	79	592	339	108	-172	1
1996	411	21	68	545	302	83	-174	1
1997	420	34	90	549	367	133	-86	5
1998	418	51	84	447	391	133	-55	3
1999	933	98	109	260	196	87	150	3
2000	991	137	93	5	0	0	201	147
2001	995	197	89	-154	170	61	203	533
2002	992	209	97	-203	288	112	89	296
2003	999	196	145	-211	194	113	-414	8

Références

Deroubaix J.-Cl. (1998). Deux langues pour une même politique : étude d'un corpus bilingue parallèle de textes politiques. In *Actes des JADT 1998.*

Deroubaix J.-Cl. (1997). *Les déclarations gouvernementales en Belgique (1944-1992).* Étude de lexicométrie politique. Thèse en sciences du langage, Sorbonne nouvelle Paris 3.

Friot B. (2003). Resource regime reforms and worker status. In Clasquin B. et Moncel N. (Eds), *Social Rights over Financial Resources : Issues for the Future of Employment and Social Protection in Europe*, Publication of the TSER European Network "Social Construction of Employment". Editions PIE-Peter Lang.

Gobin C. et Deroubaix J.-Cl. (1989). Les temps sociaux et le discours politique. Repérage de la notion de temps dans les Déclarations gouvernementales belges. *Histoire et Mesure*, vol. (3-4). Éd. du CNRS : 147-171.

Gobin C., Coron G. et Dufresne A. (2003). The European Union and resources restructuration : employment, pension and wage. In Clasquin B. et Moncel N. (Eds), *Social Rights over Financial Resources : Issues for the Future of Employment and Social Protection in Europe*. Publication of the TSER European Network "Social Construction of Employment". Editions PIE-Peter Lang.

Gobin C. (2003). L'Union européenne : l'institution politique est évanescente, le syndicat est un partenaire, le travailleur un problème, où est passé l'acteur ? Communication au colloque *The Economic's Representation of Actor at Work*. Université des Sciences et Techniques de Lille.

Kraif O. (2001). *Constitution et exploitation de bi-textes pour l'Aide à la traduction.* Thèse en sciences du langage. Université de Nice.

Martinez W. et Zimina M. (2002). Utilisation de la méthode des cooccurrences pour l'alignement des mots de textes bilingues. In *Actes des JADT 2002.*

Lebart L. et Salem A. (1994). *Statistique textuelle*. Dunod.

Génération de corpus multilingues dans la mise en œuvre d'un outil en ligne d'aide à la lecture de textes en langue étrangère

Guy Deville[1], Laurence Dumortier[1], Hans Paulussen[2]

[1]Facultés Universitaires N.D. de la Paix – 5000 Namur – Belgique
Guy.Deville@fundp.ac.be, Laurence.Dumortier@fundp.ac.be

[2]K.U. Leuven – Campus Kortrijk (KULAK) – 8500 Kortrijk – Belgique
Hans.Paulussen@kulak.ac.be

Abstract

This paper presents a method for the automatic generation of aligned bilingual corpora in a Web-based reading tool for Dutch texts by French speaking learners (NEDERLEX). The authors first discuss the major functions of NEDERLEX. Then they describe the role of bilingual corpora in the design and construction of the NEDERLEX tool, as well as the approach adopted for the extraction and alignment of such corpora. A demo of the NEDERLEX prototype will be presented during the conference talk.

Résumé

Cet article expose une méthode de génération automatique de corpus bilingues alignés dans la mise en œuvre d'un outil en ligne d'aide à la lecture de textes néerlandais à l'usage des apprenants francophones (NEDERLEX). Les auteurs présentent d'abord les principales fonctionnalités de l'outil NEDERLEX. Ils décrivent ensuite le rôle spécifique des corpus bilingues dans la construction et l'utilisation d'un tel outil, ainsi que la méthode d'acquisition et d'alignement de ces corpus. Une démonstration de la version prototype de NEDERLEX est prévue lors de l'exposé oral durant la conférence.

Mots-clés : alignement de corpus multilingues, logiciel en ligne d'apprentissage des langues étrangères.

1. Introduction

D'expérience, on observe que l'opacité du vocabulaire néerlandais et sa structure morpho-lexicale très éloignée du français constituent une des principales pierres d'achoppement dans la maîtrise de cette langue étrangère par les francophones. L'apprentissage du vocabulaire constitue dès lors une dimension importante des méthodes d'enseignement du néerlandais en tant que seconde langue, qui sont proposées sur le marché en général, et dans l'enseignement supérieur en particulier.

Les manuels traditionnels d'enseignement des langues étrangères sont principalement constitués de textes accompagnés de leur vocabulaire. Ce vocabulaire résulte d'un choix de l'auteur, et est souvent présenté en listes bilingues de mots isolés ou repris dans un contexte minimal. Les limites d'une telle présentation statique sont évidentes : (i) pour déchiffrer un texte, l'apprenant doit constamment passer « physiquement » du texte support à la liste de vocabulaire et vice-versa, (ii) le vocabulaire proposé ne couvre qu'un sous-ensemble des mots du texte ; (iii) la liste ne peut varier selon le niveau de connaissance de l'apprenant.

Les versions logicielles des cours de langues (en ligne sur la Toile ou sous forme de CD-ROM) permettent d'intégrer un dictionnaire traductif sous format électronique, consultable à la demande de l'étudiant.

Cette approche supprime les inconvénients liés au support écrit : elle offre l'accès à un glossaire pour une très large couverture lexicale des textes, et la fréquence de consultation du dictionnaire est en fonction du niveau de l'étudiant. Toutefois, le dictionnaire électronique est un outil non intégré : il est toujours consulté « hors contexte » car le choix de la traduction d'un mot dans son contexte original est laissé à l'initiative de l'apprenant, qui doit sélectionner dans le dictionnaire les informations appropriées (catégorie grammaticale, sens, exemples). Une telle lecture contextualisée de la langue est une démarche malaisée qui n'est pas à la portée immédiate de la plupart des apprenants.

Ce constat a amené les auteurs à concevoir NEDERLEX, qui est un outil original de création de supports en ligne d'aide à la lecture de textes néerlandais (Deville et Dumortier, 2003). Dans sa conception, NEDERLEX fait appel à de grands corpus bilingues alignés pour illustrer en contexte et sans ambiguïté le vocabulaire de tels textes.

Cet article est centré sur les aspects de méthodologie et d'ingénierie linguistique qui sous-tendent l'élaboration de tels corpus bilingues. Les auteurs présentent d'abord les principales fonctionnalités de l'outil NEDERLEX. Ils décrivent ensuite le rôle spécifique des corpus bilingues dans la construction et l'utilisation d'un tel outil, ainsi que la méthode d'acquisition, de génération et d'alignement de ces corpus. Ils concluent en exposant les principaux bénéfices de la génération et de l'utilisation de corpus multilingues dans NEDERLEX tant du point de vue des concepteurs de l'outil (enseignants), que de leurs utilisateurs finaux (apprenants).

2. Fonctionnalités de l'outil NEDERLEX

NEDERLEX est conçu comme un outil générique interactif (à l'usage des enseignants) pour l'édition d'un cours de textes néerlandais multimédia de tous niveaux. Cet outil génère un produit fini (à l'usage des apprenants) sous la forme d'un site Web qui offre un ensemble de documents écrits authentiques de tous types et de niveaux de difficulté différents, qui sont entièrement glosés (avec explication du vocabulaire), et assortis d'exercices interactifs. Pour chacun des textes, l'outil reprend la traduction en contexte de chaque mot, qui fait l'objet d'une série d'illustrations sous la forme de citations bilingues (concordances). Ces illustrations sont extraites d'un grand volume de textes néerlandais-français alignés, appelés ici « corpus bilingues alignés ». D'un point de vue technique, le site Web est généré par des pages dynamiques construites en PHP via une interaction avec une base de données MySql reprenant sous forme de table les ressources linguistiques associées à chaque cours (textes des leçons, lexique, homonymes, concordances et corpus bilingues alignés).

NEDERLEX se présente sous la forme d'une interface reprenant les fonctionnalités nécessaires à la création d'un cours multimédia, à savoir :

(i) créer et éditer une leçon à partir d'un fichier texte. Cette leçon est ensuite balisée : pour chaque mot de la leçon, on détermine de manière automatique sa clé de la table lexique (associant des informations telles que lemmes et formes fléchies néerlandaise et française, catégorie syntaxique, genre, etc.) et on associe à ce mot une fonction javascript qui permet d'afficher, au clic de souris, toutes les concordances trouvées dans les corpus alignés (stockées dans la table des concordances) qui correspondent à ce lemme. Une fonction sous forme de menu déroulant permet à l'utilisateur de lever les ambiguïtés qui perturbent la traduction de certains mots, en raison de phénomènes de polysémie et d'homonymie ; Ce menu déroulant présente à cet effet les informations issues de la table des homonymes.

(ii) importer, mettre à jour et contrôler le lexique associé au texte d'une leçon (le lexique du prototype compte actuellement environ 4.000 entrées validées et exploitées dans le cadre de cours existants, sur un total d'environ 16.000) ;

(iii) importer et mettre à jour la base de corpus bilingues alignés qui produiront les extraits illustrant chaque mot des leçons du cours et sa traduction en contexte (concordances). L'élaboration de ces corpus (extraction et alignement par phrase) est réalisée de manière semi-automatique. Le jeu de corpus alignés a un volume total de 1.500.000 mots ;

(iv) générer l'ensemble des concordances, c'est-à-dire les extraits bilingues des corpus alignés dans lesquels les mots du lexique et leur traduction apparaissent ; A cet effet, chaque entrée de la table des concordances contient la clé du lemme, ses formes fléchies néerlandaise et française apparaissant dans le corpus aligné, ainsi que les numéros de référence du corpus et du paragraphe concernés.

Pour gérer toutes ces fonctionnalités, nous disposons de cinq tables reprises dans la base de donnée décrite ci-dessus : la table des textes, la table du lexique, la table des homonymes, la table des corpus bilingues alignés et la table des concordances.

La méthodologie retenue favorise un développement et une mise à jour modulaires — et en grande partie automatisés — des textes de leçons, des lexiques associés et du corpus de textes bilingues alignés, selon le schéma de la figure 1.

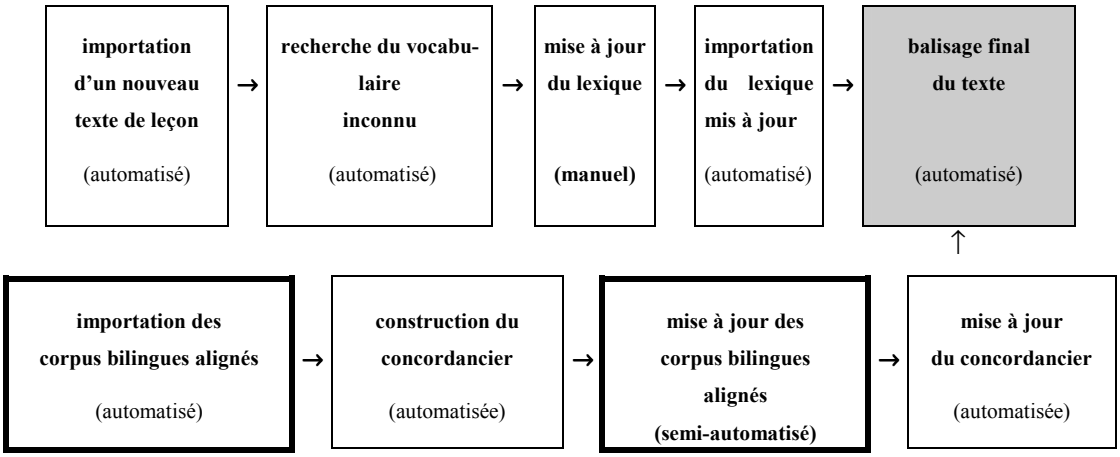


Figure 1. Schéma de développement d'un cours avec l'outil NEDERLEX

L'outil NEDERLEX génère donc un texte de lecture qui a été préalablement balisé, ce qui permet à l'apprenant d'« interroger » chaque mot d'un clic de souris. Lorsque l'apprenant clique sur un mot du texte pour en solliciter la traduction, un tableau apparaît dans une fenêtre en bas de l'écran, avec les informations suivantes : (i) la ligne supérieure affiche les informations grammaticales du mot : forme lemmatisée, catégorie et sous-catégorie syntaxique et formes fléchies pertinentes d'un point de vue didactique (pluriel des noms, forme comparative et superlative des adjectifs, formes prétérit et participe passé des verbes) ; (ii) les cellules inférieures du tableau reprennent sous leur différentes formes fléchies, plusieurs occurrences du mot néerlandais (cellules gauches) avec leurs différentes traductions françaises (cellules droites) dans des contextes ou « concordances » issus de corpus bilingues alignés.

Par un clic sur le mot néerlandais dans son contexte (cellule de gauche), on ouvre une nouvelle fenêtre qui affiche le contexte complet du mot retenu (phrase en néerlandais avec sa traduction en français), avec l'indication de la référence de cette source. Ces fenêtres sont de

couleurs différentes en fonction du type de corpus d'où proviennent les concordances (voir section 3.). Les entités balisées peuvent être des mots simples ou composés ainsi que des unités lexicales constituées de plusieurs mots (formes disjointes de verbes séparables, locutions, syntagmes verbaux ou nominaux, etc.).

On notera que les exemples de traduction de chaque mot sont toujours donnés en fonction du contexte de ce mot dans la leçon, les ambiguïtés ayant été levées lors du balisage préalable du texte ; il y a donc absence de polysémie et d'homonymie. Ce choix désambiguïsé et contextualisé constitue une aide précieuse pour l'apprenant, qui est guidé dans sa recherche de la traduction exacte des mots inconnus. La figure 2. reprend un exemple de texte de lecture développé avec l'outil NEDERLEX.

La structure optimisée des tables permet un affichage rapide des informations lexicales sollicitées par l'utilisateur. En effet, chaque clic de souris ne requiert l'accès qu'à une seule table : soit la table des concordances (en cas de clic sur un mot de la leçon pour faire apparaître les concordances), soit la table des corpus bilingues alignés (en cas de clic sur un mot du tableau des concordances pour faire apparaître le contexte complet du mot retenu).

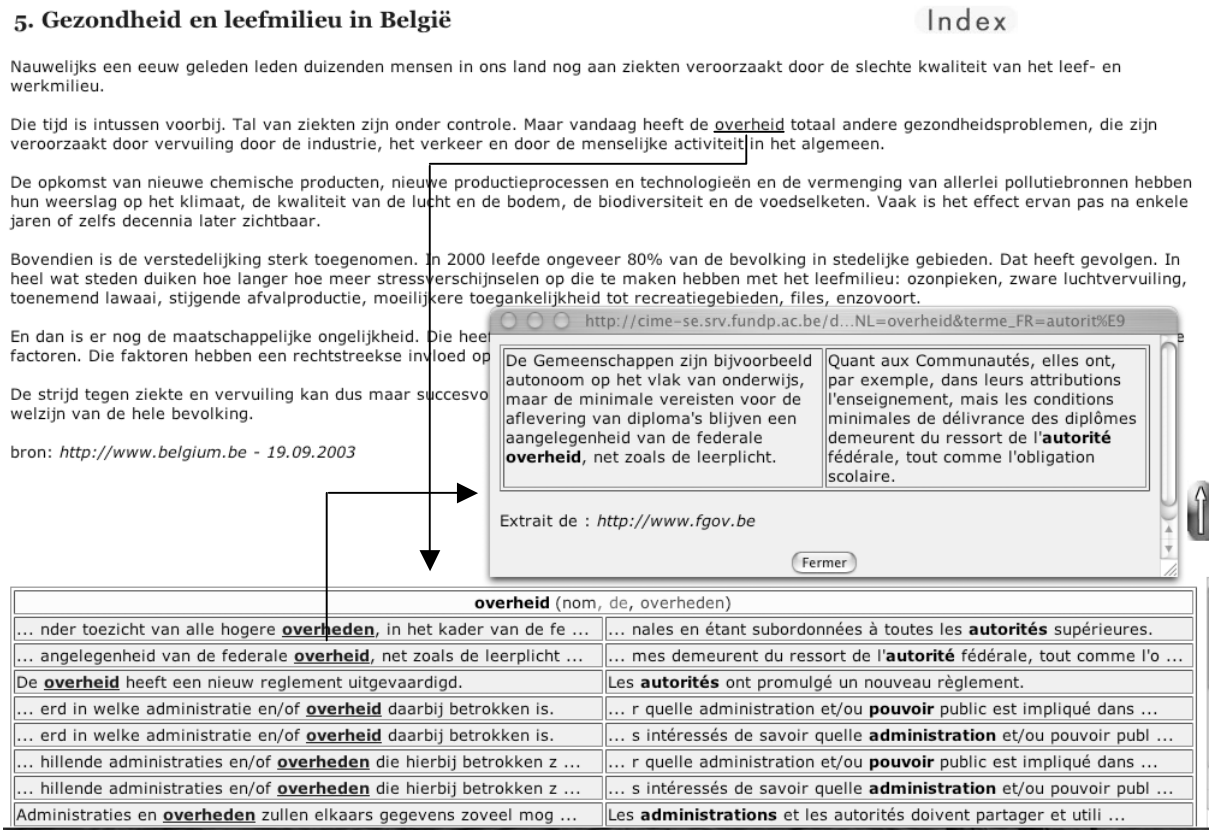


Figure 2. Extrait d'un cours développé avec l'outil NEDERLEX

3. Génération semi-automatique de corpus bilingues alignés

Les corpus bilingues néerlandais-français alignés constituent donc une des ressources linguistiques principales de NEDERLEX. Ces corpus sont de trois types : (i) les supports écrits de cours de néerlandais que nous avons édités jusqu'à présent comportent de précieux glossaires éprouvés sur le plan didactique, qui reprennent plusieurs milliers de phrases avec leur traduc-

tion (néerlandais-français) ; nous avons reformaté ces phrases sous forme de corpus aligné ; (ii) plusieurs sites Web présentent aujourd'hui de nombreux textes de grande qualité dans plusieurs versions linguistiques (dont le français et le néerlandais). Il s'agit de sites fédéraux officiels (gouvernement belge, ministères, cours et tribunaux), de sites d'organismes internationaux (par ex. l'Union européenne) ou encore de sites commerciaux (par ex. dans le secteur de la distribution alimentaire) ; (iii) enfin plusieurs sites présentent des textes de nature technique (par ex. des textes de loi, jugements et jurisprudence) qui nous servent à illustrer un cours de terminologie juridique néerlandaise (Deville et Dumortier, 2002).

Les corpus de type (i) sont constitués manuellement, car ils possèdent d'emblée une structure proche du format souhaité (alignement au niveau de la phrase), et constituent un corpus fini, stable et homogène. Dans une première phase, les corpus de type (ii) et (iii) ont été constitués manuellement en identifiant des versions néerlandaise et française de textes pertinents tels que décrits plus haut, qui sont copiées dans un tableur et alignées ensuite au niveau du paragraphe. Cette procédure fastidieuse a un double inconvénient : la mise à jour et l'enrichissement systématique des corpus est malaisée et les corpus ne sont alignés qu'au niveau du paragraphe. Dans le cas de longs paragraphes, cette dernière contrainte rend la lecture du mot-clé et de sa traduction peu lisible, et il arrive que ce mot-clé soit restitué dans une traduction erronée, pour des raisons techniques liées à l'algorithme de construction des concordances. Pour ces motifs, nous avons décidé d'automatiser (i) l'élaboration de corpus à partir de sites Web multilingues tels que décrits plus haut, et (ii) l'alignement de ces corpus au niveau de la phrase au lieu du paragraphe. Détaillons à présent cette démarche.

Le *Corpus Namur* développé dans le cadre d'une thèse de doctorat (Paulussen, 1999) constitue un exemple de corpus multilingues qui a été créé de manière semi-automatique[1]. Il contient des textes en français, anglais et néerlandais d'un volume total de 2.000.000 mots. Une moitié du corpus contient des textes littéraires (fiction), l'autre moitié contient des textes autres que de fiction : les débats du Parlement européen d'une part, et un volume du Courier de l'Unesco de l'autre. L'utilisation d'outils développés en Awk et Perl, nous a permis de nettoyer et d'aligner au niveau du paragraphe les textes multilingues du *Corpus Namur* en provenance de différentes plates-formes informatiques.

Au début des années 90, la standardisation de l'encodage de tels textes était à un stade embryonnaire, et ceux-ci ne comportaient aucune forme d'annotation (tels les balises HTML), de sorte qu'un texte était simplement défini comme un bloc de lignes (d'une longueur d'environ 60 caractères), chaque ligne se terminant par une fin de ligne, et un bloc se terminant par deux lignes. Grâce à une procédure semi-automatique d'alignement, une sélection de phrases échantillons a été automatiquement convertie en une base de données trilingues (TRIPTIC), qui constituait la plate-forme de travail pour une étude en linguistique contrastive que nous ne détaillerons pas ici.

Cette approche réalisée dans un environnement Unix/Linux s'est avérée très puissante, mais limitée à des utilisateurs quelque peu spécialisés de la programmation. Une nouvelle approche consiste à présent à raffiner et intégrer dans un environnement convivial les outils que nous avons préalablement créés, afin de permettre à tout utilisateur d'exécuter les différentes étapes de sélection et d'alignement de corpus multilingues par un simple clic de souris. Cette approche est réalisable aujourd'hui grâce au développement de l'internet et à la standardisation des formats de textes.

[1] Voir également : http://www.fundp.ac.be/~hpaulus/NamurCorpus.html

Depuis l'avènement et la diffusion des navigateurs (browsers) HTML dans les années 90, tout internaute peut aujourd'hui récupérer n'importe quel fichier du monde entier en provenance de tout type de plate-forme informatique, sans devoir maîtriser les commandes énigmatiques d'outils spécialisés de programmation (tels que *ftp*, par exemple). La page Web ou le fichier à récupérer se trouve à la portée d'un simple clic de souris. Chaque clic active une connexion selon le protocole HTTP, dont les détails se déroulent dans les coulisses du navigateur. La consultation de pages HTML s'opère la plupart du temps selon ce mode, c'est-à-dire un simple clic sur un lien hypertexte dans une page HTML affichée dans la fenêtre d'un navigateur. On peut également consulter tout serveur Web (qui contient des documents HTML) à l'aide de n'importe quel langage de programmation ou langage script qui supporte le protocole HTTP. Cette consultation non interactive permet d'automatiser le processus de récupération régulière de fichiers HTML, ce qui peut s'avérer très utile lorsqu'on visite fréquemment des sites « périodiques », qui changent à un rythme quotidien, hebdomadaire ou mensuel.

Aussi simple que cette approche puisse paraître, il n'empêche que la démarche exige un travail de programmation défensive de la part des développeurs de tels outils, car la connexion internet n'est pas toujours fiable et la structure du site à consulter peut changer d'un jour à l'autre. Heureusement, la gestion des sites Web se stabilise de manière croissante, puisqu'une diffusion de qualité des données exige une structure plus rigide permettant une mise à jour efficace des informations disponibles. Ce constat est particulièrement vrai dans le cas de sites multilingues. Ainsi, la stabilisation de la structure des sites Web et les performances accrues des fonctionnalités HTTP dans les langages script nous ont permis d'envisager la récolte automatique de textes multilingues en ligne. Cette récolte est la première étape dans la construction d'un module de génération automatique de textes multilingues. Les étapes suivantes comportent une procédure qui divise les textes en unités de phrases et une procédure automatique d'alignement de ces phrases.

L'alignement automatique de textes au niveau de la phrase est une condition préalable à une exploitation efficace d'un corpus parallèle. Un tel alignement implique la mise en correspondance automatique des portions sélectionnées d'un texte source et des portions équivalentes dans un texte cible. Il existe aujourd'hui plusieurs outils d'alignement, mais la plupart d'entre eux requièrent un imposant travail de post-édition, que nous voulions réduire de manière significative. En outre, aucun outil d'alignement n'a été développé spécifiquement pour le néerlandais.

Les algorithmes d'alignement utilisent tantôt une approche statistique tantôt une approche lexicale, bien qu'un mélange des deux approches soit de plus en plus utilisé (Brown *et al.*, 1990 ; Gale et Church, 1991 ; Simard *et al.*, 1992 ; Church, 1993 ; McEnery *et al.*, 1997).

(Simard *et al.*, 1992) sont probablement les premiers à allier une approche statistique avec un support linguistique, en introduisant la notion de « cognates ». Leur approche est une amélioration du modèle probabiliste de (Gale et Church, 1991) qui est strictement basé sur la longueur de phrase. Une approche plus linguistique est utilisée dans le programme d'alignement développé par (Hofland, 1996), dans le cadre du projet ENPC (English-Norwegian Parallel Corpus). Dans cette étude, Hofland utilise des mots appelés « ancres », qui sont stockés dans un lexique bilingue contenant des mots qui sont soit (i) raisonnablement fréquents, ou (ii) qui ont des équivalents transparents dans chacune des deux langues utilisées.

Dans le cadre de cet article, nous présentons une série préliminaire de tests d'alignement de corpus bilingues selon une procédure qui comporte les étapes suivantes :

La première étape consiste à télécharger automatiquement un ensemble de textes parallèles (néerlandais-français) de bonne qualité à partir de sites web sélectionnés sur base de leur structure bilingue, comme indiqué plus haut. Notre choix s'est porté sur (i) des textes relatifs à l'alimentation repris sur le site d'une chaîne de grande distribution (www.delhaize.be), et sur (ii) la transcription des débats parlementaires européens accessibles sur le site du Parlement européen (www.europarl.eu.int/plenary). En outre, (iii) le texte du dernier accord gouvernemental a été repris manuellement du site du gouvernement fédéral belge (www.belgium.be) à partir d'un fichier pdf. Le volume de chaque échantillon de textes est repris au tableau ci-dessous.

Le nettoyage de ces fichiers HTML en simples fichiers textes a été réalisé à l'aide d'un navigateur texte (lynx) qui permet d'extraire automatiquement les balises d'un fichier HTML. Cette approche, qui exige un nettoyage supplémentaire du texte à l'aide de filtres (scripts) écrits en langages Awk et Perl, est perfectible. Ainsi, à l'avenir, cette procédure devrait être optimisée par l'utilisation de parseurs de documents HTML, dont des versions efficaces pour Perl ou PHP sont librement disponibles.

Nous avons ensuite écrit une procédure d'identification automatique des phrases en Perl (splitSentence.pl) qui s'applique sur ces textes nettoyés. Pour ce faire, nous avons utilisé le module Sentence.pm, développé par Shlomo Yona (http://search.cpan.org/~shlomoy), et qui est téléchargeable à partir du site *Comprehensive Perl Archive Network* (www.cpan.org).

L'alignement des phrases de chaque version linguistique des textes (néerlandais et français) a été réalisé à l'aide du programme d'alignement développé par (Danielsson et Ridings, 1997), qui est entièrement basé sur l'algorithme de (Gale et Church, 1991). Cet algorithme a l'avantage de fonctionner indépendamment de la langue des textes à aligner.

TEXTE	TAILLE (# mots)	SCORE (%) mode automatique	SCORE (%) mode semi-automatique
DELHAIZE	13 688	83,78	94,17
ACCORD GOUVERNEMENTAL	51 332	78,30	96,77
DEBATS PARLEMENT EUROPEEN	46 635	88,93	92,44
TOTAL	**111 655**	**83,67**	**94,46**

Résultats de l'algorithme d'alignement
de trois échantillons de textes néerlandais-français

Cette série préliminaire de tests nous a amenés à faire les observations suivantes : tout d'abord, nous avons constaté que les sites web multilingues dont nous avons extrait les textes sont structurés de manière très variée. Ainsi, le site www.delhaize.be (textes relatifs à l'alimentation) est une arborescence de niveau variable constituée à la fois de pages HTML intermédiaires (branches) comprenant des hyperliens et de pages HTML finales (feuilles) exemptes d'hyperliens. Cette structure particulière a nécessité l'écriture d'un script sur mesure permettant l'extraction des fichiers textes néerlandais et français du site. Inversement, le site www.europarl.eu.int/plenary (débats du Parlement européen) est très strictement structuré et hiérarchisé, ce qui nous a permis d'écrire un script générique qui extrait les transcriptions des débats parlementaires dans les langues souhaitées (néerlandais-français). L'architecture rigoureuse d'un tel site est motivée par la nécessité d'une mise à jour optimisée car une très grande quantité de fichiers vient l'enrichir à une fréquence hebdomadaire.

Comme indiqué dans le tableau, l'outil d'alignement a été appliqué sur les textes selon deux modes : (i) dans le mode automatique, les textes ont été extraits du site web, nettoyés, découpés en phrases et alignés sans aucune intervention manuelle ; (ii) dans le mode semi-automatique, l'identification (ou découpage) en phrases a été corrigé manuellement avant l'alignement automatique proprement dit. Le score de l'algorithme exprime (en %) le quotient du nombre d'alignements corrects par le nombre total d'alignements générés automatiquement.

Dans le mode automatique, l'algorithme d'alignement obtient des scores plutôt faibles pour les textes moins structurés — qu'il soient sous forme de pages web ou en format pdf — (83,78 % pour les textes *delhaize* et 78,30 % pour les textes *accord gouvernemental*), alors que ce score s'élève à 88,93 % pour les textes *débats du Parlement européen*, à la structure plus rigoureuse.

Dans le mode manuel, l'algorithme d'alignement obtient des scores sensiblement supérieurs au mode automatique pour ces même textes moins structurés (94,17 % pour les textes *delhaize* et 96,77 % pour les textes *accord gouvernemental).* Cette nette amélioration des résultats par rapport au mode automatique s'explique par des différences typographiques importantes des versions néerlandaise et française de ce type de textes, alors qu'on note une moindre amélioration (92,44 %) dans le cas des textes *débats du Parlement européen*, qui observent une structure typographique nettement plus normalisée.

D'une part, ces résultats plaident en faveur d'une amélioration de l'algorithme de segmentation des textes en phrases, notamment lors du traitement de certains signes de ponctuation tels que les suites de caractères «). », « ... mot », « ...) », « - mot » qui sont incorrectement interprétés dans certains cas.

Ensuite, nous devrions procéder à une comparaison des performances de notre programme d'alignement avec d'autres outils existants, ce qui sort du cadre strict de cet article. Enfin, nous comptons tester une approche plus linguistique en vue d'améliorer les performances de notre algorithme d'alignement automatique de corpus bilingues.

4. Conclusion

Nous avons présenté NEDERLEX, un outil original de création de supports en ligne d'aide à la lecture de textes néerlandais, qui exploite de grands corpus bilingues alignés (néerlandais-français) pour illustrer en contexte et sans ambiguïté le vocabulaire de tels textes.

Lors de la mise en œuvre de cet outil, nous avons développé une méthode d'acquisition, de génération et d'alignement de ces corpus bilingues au niveau de la phrase. Les premiers résultats de l'algorithme d'alignement appliqué sur des échantillons représentatifs de notre base de donnée de corpus bilingues — constituée manuellement — sont encore timides mais encourageants : ils génèrent en moyenne plus de 94 % d'alignements corrects (avec une correction manuelle des textes après la phase de découpage en phrases). L'objectif de réduire de manière significative le travail de post-édition manuelle a donc été rencontré. Un des points perfectibles de l'algorithme réside dans l'utilisation de connaissances linguistiques lors de la phase d'alignement.

Sans insister sur les avantages évidents d'une aide lexicale en ligne intégrée dans un outil générique de lecture de textes en langue étrangère sur le Web, nous mentionnerons les principaux bénéfices de l'utilisation de corpus multilingues dans un tel outil.

Du point de vue des concepteurs de l'outil (enseignants), l'acquisition systématique de très grands corpus bilingues a été facilitée par une série d'outils d'extraction de textes multi-

lingues sur le Web et d'alignement de ces textes, qui optimisent le travail de post-édition. Notons que les outils mentionnés n'exigent pas de compétence informatique particulière de la part des utilisateurs.

Du point de vue des utilisateurs finaux (apprenants), les mots des leçons sont systématiquement illustrés à l'aide d'exemples issus de ces corpus bilingues alignés. L'apprenant peut ainsi observer le mot recherché sous plusieurs formes fléchies et sa traduction dans plusieurs contextes minimaux et différenciés, qui sont spécifiques au contexte du mot candidat. Cette aide lexicale désambiguïsée et contextualisée constitue une des plus fortes valeurs ajoutées du système. Notons également que tous les mots des textes font systématiquement l'objet d'une telle traduction, ce qui rend l'outil fortement adaptatif : un même texte peut être déchiffré par des apprenants de niveaux différents.

Références

Brown Peter F., Cocke J., Della Pietra S., Della Pietra V., Jelinek F., Lafferty J., Mercer R. et Roossin P. (1990). A statistical approach to machine translation. *Computational Linguistics*, vol. (16/2) : 79-85.

Church K. (1993). Chat_align: a program for aligning parallel texts at the character level. In *Proceedings of the 31st annual meeting of the AC*L : 1-8.

Danielsson P. et Ridings D. (1997). Practical presentation of a vanilla aligner. Paper presented at the telri workshop in alignment and exploitation of texts, llubljana, feb. 1-2 1997. Research reports from the Department of Swedish, Göteborg University GU-ISS-97-2, Språkdata.

Deville G. et Dumortier L. (2002). Tussen de Regels – deel II, Lecture de textes juridiques néerlandais, cours en ligne (www.droit.fundp.ac.be/langues/termino_nl.htm). Facultés universitaires de Namur.

Deville G. et Dumortier L. (2003). Tussen de Regels – deel I, Lecture de textes néerlandais, cours en ligne (www.droit.fundp.ac.be/langues/nl.htm). Facultés universitaires de Namur.

Gale W. et Church K. (1991). A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics* : 177-184.

Hofland K. (1996). A program for aligning English and Norwegian sentences. In G. Perissinotto (Ed.), *Research in Humanities Computing*, vol. (5). Oxford University Press : 165-178.

McEnery A, Oakes M. et Garside R. (1997). CRATER: resource creation for corpus-based machine translation. In Lewandowska-Tomaszczyk B. et Thelen M. (Eds), *Translation and meaning*, Part (4) : 495-500.

Paulussen H. (1999). *A corpus-based contrastive analysis of English 'on/up', Dutch 'op' and French 'sur' within a cognitive framework.* Thèse de doctorat non publiée, Université de Gand.

Simard M., Foster G. et Isabelle P. (1992). Using cognates to align sentences in bilingual corpora. In *Proceedings of the fourth international conference on theoretical and methodological issues in machine translation* (TMI92) : 67-81.

La féminisation des noms de métier, fonction, grade ou titre en Belgique francophone. État des lieux dans un corpus de presse

Anne Dister

Cental et Centre de recherche VALIBEL – UCL
Louvain-la-Neuve – Belgique
dister@tedm.ucl.ac.be

Résumé

En 1993, suivant en cela d'autres pays francophones, le Conseil de la Communauté française de Belgique adopte un décret visant à féminiser les noms de métier, fonction, grade et titre. Les règles de féminisation, proposées par le Conseil supérieur de la langue française (de Belgique), doivent être appliquées dans tous les textes administratifs ainsi que dans toutes les publications qui émanent d'institutions et d'associations subventionnées par la Communauté française.

Dans cet article, nous analyserons certains noms de profession dans la presse, sur base d'un corpus de 6 millions de mots issus des quatre derniers mois de l'année 2001 du journal belge *Le Soir*. En nous centrant particulièrement sur certaines formes, sélectionnées parce qu'emblématiques des réticences à la féminisation, nous verrons où en est la Belgique francophone dans la reconnaissance, à travers les pratiques linguistiques, des femmes au travail.

Mots-clés : féminisation, nom de métier, nom générique, nom spécifique, référent, politique linguistique, Communauté française de Belgique, linguistique de corpus.

1. La féminisation en Belgique francophone

Le 21 juin 1993, le Conseil de la Communauté française de Belgique adopte un décret visant à féminiser les noms de métier, fonction, grade et titre. L'article 1er du décret précise que les règles de féminisation devront être appliquées dans tous les textes administratifs ainsi que dans toutes les publications qui émanent d'institutions et d'associations subventionnées par la Communauté française.

Le 5 juillet 1993, sous la houlette de son président Jean-Marie Klinkenberg, le Conseil supérieur de la langue française, l'organe consultatif chargé de conseiller le ministre ayant la langue dans ses attributions, rend son avis ; le 13 décembre 1993, le Gouvernement de la Communauté française prend un arrêté définissant les règles de féminisation, arrêté d'application qui suit les recommandations du Conseil supérieur de la langue française.

Ainsi, après le Québec, pionnier en la matière pour la francophonie (la *Gazette officielle* donne des recommandations aux administrations dès 1979), la France en 1986 (avec la circulaire Roudy qui, il est vrai, ne sera jamais appliquée) et la Suisse (le canton de Genève adopte une loi dès 1988, qui préconise l'utilisation des formes féminisées dans l'administration), la Belgique francophone se dote d'un décret en ce qui concerne l'utilisation, dans les textes de

ses administrations et de ses institutions, de formes féminisées pour désigner les femmes au travail.

2. Pourquoi une intervention de l'État en matière de langue ?

Depuis la fin de la seconde guerre mondiale, moment où elles obtiennent le droit de vote, les femmes exercent des professions de plus en plus variées et accèdent à des postes à responsabilités ou à des fonctions dites prestigieuses. Or, cette présence grandissante des femmes dans le monde du travail, à tous ses échelons, ne se reflète pas dans les termes qui les désignent : on dit d'une femme qu'elle est *chirurgien*, *chercheur*, *ambassadeur* ou encore *huissier de justice*. Les femmes sont maintenant partout, mais le nom de leur profession reste au masculin. Comme le rappelle Marie-Louise Moreau (1999), « Au moyen âge, à la Renaissance, à l'époque classique, au 18e, au 19e, les francophones ont systématiquement utilisé des termes au féminin pour désigner les femmes », et de rappeler les *drapières*, *tisserandes*, *abbesses* ou encore la *régente* Marie de Médicis.

Se pose alors la question suivante : pourquoi une telle réticence, aujourd'hui, à parler de *la chirurgienne*, de *la chercheuse*, de *l'ambassadrice* ou de *la huissière de justice* lorsque l'on a affaire à des femmes ? Et pourquoi cette réticence se rencontre-t-elle bien souvent chez les femmes elles-mêmes, qui répugnent à utiliser la forme féminine et préfèrent le masculin ? Tout se passe comme si, pour être l'équivalent des hommes, ces femmes qui accèdent à des postes de prestige devaient recevoir le même titre que les hommes, sans distinction morphologique. Tout le monde trouve banal de dire *une secrétaire*, mais lorsqu'elle accède à un poste important à l'Académie française, Hélène Carrère d'Encausse exige de se faire appeler *Madame le secrétaire perpétuel*. Lorsque la fonction est prestigieuse, elle devrait être déclinée au masculin. Perdrait-elle (la fonction, et la femme qui l'occupe) de son prestige au féminin ?

D'autre part, certains noms de professions considérées comme non prestigieuses montrent également des réticences à la féminisation. Des arguments qui relèvent plus de la psycholinguistique que de la sociolinguistique entrent en jeu : certaines formes seraient vulgaires (*sauteuse*, *entraineuse*, etc.), d'autres, peu élégantes, voire barbares, sonneraient terriblement mal (*écrivaine*, *rapporteuse*), d'autres encore seraient grandement ambiguës (c'est le cas de la fameuse *cafetière* sur laquelle l'attention des détracteurs du décret s'est focalisée).

La volonté déclarée des auteurs du décret était d'assurer la visibilité des femmes à travers la langue. En effet, l'utilisation d'une seule forme, au masculin, pour désigner les hommes et les femmes dans l'exercice de leur profession occulte la place qu'occupent aujourd'hui les femmes dans le monde du travail.

Les détracteurs à la féminisation ont avancé divers arguments[1], dont notamment la fonction de neutre du masculin. Le masculin serait le genre non marqué tandis que le féminin serait le genre marqué. Le masculin servirait donc aussi bien à désigner l'homme que la femme. Or, on sait à quel point « dans nos usages et nos représentations, le genre masculin est associé au trait mâle » (Houdebine, 1987 : 19). L'anecdote célèbre du chirurgien qui ne peut opérer son fils, blessé lors d'un accident de voiture dans lequel son père a trouvé la mort, est révélatrice à cet égard. Qui pense d'emblée que ce chirurgien est une femme, la mère de l'enfant ?

Ce qui n'est pas nommé publiquement n'existe pas socialement. D'où la nécessité d'une politique de la langue qui rende visible la place des femmes dans notre société.

[1] « La typologie des attaques mérite d'être signalée tant elle est pauvre : politique, sexiste, prescriptive » (Houdebine, 1987).

3. Le corpus

Notre corpus est constitué de quatre mois d'articles du journal *Le Soir* (dernier trimestre de 2001)[2]. Ce corpus, de 37 mégas, comprend près de 6 millions de mots[3].

Nous avons contacté la rédaction du *Soir* afin de savoir si des consignes avaient été données aux journalistes en ce qui concerne la féminisation des noms de profession. Le chef de la qualité du journal nous a dit que certains journalistes étaient plus sensibilisés que d'autres au problème, mais que le libre-arbitre était laissé à chacun. « Mais on essaie d'écrire *la* ministre ou encore l'*auteure* avec un *e.* » Nous verrons concrètement, à travers certains exemples, ce qu'il en est.

4. La méthodologie

Notre objectif, dans cet article, est d'observer les pratiques effectives dans la presse écrite en ce qui concerne l'utilisation de formes féminisées de noms de profession pour désigner des femmes[4].

Nous avons décidé de nous pencher plus particulièrement sur certaines forment réputées rétives à la féminisation : « De nombreuses formes masculines résistent à la féminisation, soit pour des raisons morphologiques, soit par le seul poids des traditions » (Muller, 1994). Comme le rappelle Marie-Josèphe Mathieu (1999), des réticences existent, qui concernent trois grandes classes de professions : « les activités valorisées symboliquement et financièrement, celles qui sont naturelles et/ou physiques et celles dont le féminin implique une connotation sexuelle […] ».

Notre méthodologie[5] est la suivante : pour les noms de profession que nous avons décidé d'analyser, nous avons extrait systématiquement de notre corpus toutes les occurrences au masculin singulier, au féminin singulier et au féminin pluriel. Nous avons d'emblée écarté les formes au masculin pluriel qui 1) soit réfèrent à l'ensemble des personnes qui exercent la profession, sans distinction de sexe, 2) soit réfèrent à des hommes et des femmes en particulier, mais « le masculin l'emporte sur le féminin ». Parmi les formes au masculin singulier, nous avons alors regardé le contexte afin d'isoler les quatre cas de figure suivants :

1. l'emploi de la forme est dit spécifique et le référent est un homme ;
2. l'emploi de la forme est dit spécifique et le référent est une femme : dans ce cas, nous aurions attendu la forme féminisée ;
3. ni le contexte ni nos connaissances du monde ne nous permettent de trancher, mais l'occurrence réfère à un individu en particulier (emploi spécifique) ;
4. l'emploi du terme masculin est un emploi générique.

[2] *Le Soir* est, avec *La Libre Belgique*, le quotidien francophone le plus important de Belgique. Il est tiré à plus de 150 000 exemplaires (on estime le nombre de lecteurs à plus d'un demi-million, sans compter la consultation maintenant possible sur internet). Nous avons constitué notre corpus à partir du CD-Rom qui rassemble les articles de fond d'une année, en l'occurrence le dernier trimestre de l'année 2001 (il s'agit du dernier CD disponible au moment de l'écriture de cet article). Ne sont pas repris dans ce CD-Rom, notamment, les encarts publicitaires ou encore les avis nécrologiques. Ces derniers en manquent pourtant pas d'intérêt pour le sujet qui nous occupe ici.

[3] Il aurait été intéressant de faire une comparaison des pratiques avant et après le décret, mais les archives du *Soir* ne sont pas disponibles avant la date d'application du décret (1er janvier 1994). Pour une comparaison avant/après en France, voir Fujimura (à paraître).

[4] Pour une étude plus ancienne dans la presse, on se reportera à Boel (1976).

[5] Nous nous distinguons donc de Itsuko Fujimura qui, dans une étude à paraître sur le même sujet, reprend exclusivement les noms de profession dans le contexte proche d'un prénom de femme. Nous avons ici systématisé la recherche à toutes les occurrences.

5. Classes de professions

À travers certains exemples concrets, nous allons voir les pratiques en ce qui concerne les professions dont le féminin aurait un caractère sexuel, celles dont le féminin relève d'une « féminisation conjugale » ainsi que les professions dites prestigieuses (domaine de la justice, de la politique ou encore du monde de l'armée).

5.1. *Professions à connotation grivoise*

Parmi les professions dont le féminin aurait une connotation grivoise, nous avons sélectionné les formes *sauteur/sauteuse*, *coureur/coureuse* et *entraîneur/entraîneuse*.

Pour le souple *sauteur/sauteuse,* nous recueillons seulement 4 occurrences dans le corpus : 3 réfèrent à des femmes et utilisent la forme féminisée *sauteuse* (2 fois *sauteuse en longueur* et une fois *sauteuse* seule), et le seul *sauteur* réfère bien à un homme.

En ce qui concerne *coureur*, la répartition est la suivante :

	Emploi particulier	Emploi générique
coureur	77	18
coureuse	2	/

Tous les emplois particuliers de la forme masculine correspondent bien à des référents hommes[6]. Ici aussi, les 2 seules fois où l'on désigne une femme (une fois *coureuse* seule, l'autre fois *coureuse de demi-fond*), la forme au féminin est utilisée.

Notons par ailleurs que nous rencontrons également le mot dans son sens péjoratif, au masculin (4 fois) comme au féminin (2 fois) :

> L'oncle **Luden**, c'est le coureur de jupons invétéré. (Alain Delaunois, 12 déc. 2001, p. 6.)
>
> Jadis, dans nos régions, on disait d'une femme négligente ou **coureuse** que c'était une « bizouye ». (Éric Meuwissen, 8 octobre 2001, p. 15.)

Pour *entraîneur*, nous avons 1033 occurrences, qui soit réfèrent à des hommes, soit sont utilisées dans des emplois génériques. Aucune forme féminine *entraîneuse*[7], et aucun référent femme. Ceci s'explique sans doute par le fait que la grande majorité des occurrences relèvent du monde du football.

Pour les formes choisies ici, on constate que l'éventuelle connotation grivoise avancée par les détracteurs du décret n'empêche pas leur utilisation. Nous verrons en 5.3.2. ce qu'il en est de *députée*.

5.2. *Le féminin « conjugal »*

Parmi les arguments mis en avant par les opposants du décret, on trouve l'argument historico-linguistique du « féminin conjugal » : certains grades ou fonctions au féminin seraient classiquement réservés à nommer l'épouse de l'homme qui occupe la fonction. *Pharmacienne* et *ambassadrice* en sont les exemples classiques : la pharmacienne est la femme du pharmacien et madame l'ambassadrice est la femme de l'ambassadeur. Voyons comment ces termes sont utilisés dans notre corpus.

[6] Cette écrasante majorité s'explique par les nombreux articles traitant de cyclisme et de formule 1.
[7] Ni d'ailleurs de forme *entraineuse* ou *entraineuses* en orthographe réformée.

	Emploi particulier	Emploi particulier indéterminé	Emploi générique
pharmacien	8	9	8
pharmacienne	3	/	/

Les emplois spécifiques de *pharmacien* correspondent tous à des hommes ; les 3 référents qui sont des femmes sont tous désignés par le terme *pharmacienne*. Par ailleurs, en étendant notre recherche à l'ensemble de l'année 2001 du journal, nous obtenons 13 occurrences de *pharmacienne* qui correspondent toutes à des femmes qui exercent la profession de pharmacienne.

En ce qui concerne le couple *ambassadeur/ambassadrice*, nous n'avons là non plus aucune *ambassadrice* « femme de » ; par ailleurs, aucune *ambassadrice* n'exerce de fonction diplomatique. Les 13 occurrences de *ambassadrice* apparaissent dans le sens « femme chargée d'une mission » : 2 *ambassadrice de charme*, une *ambassadrice des droits de l'homme* ou encore une *ambassadrice du fado*. Notons que ce sens n'est pas inexistant pour les référents hommes : il apparaît 11 fois sur les 166 occurrences de *ambassadeur* : Pelé est *ambassadeur de l'Unicef*, Joël Smet *ambassadeur du Festival du conte de Chiny*, et Olivier Theyskens est *un excellent ambassadeur du stylisme bruxellois*.

Notons une occurrence de *ambassadeur* qui aurait pu être féminisée, même si le référent n'est pas à proprement parler un humain :

> C'est dans ce contexte que l'organisation d'Europalia a choisi la Pologne comme nouvel ambassadeur culturel. (Laurent Ancion, 3 octobre 2001, p. 5.)

En ce qui concerne la forme *colonelle*, la seule occurrence de *colonelle* recueillie dans le corpus atteste cet emploi d'un titre accordé en tant que « femme de » :

> **La veuve respectable du colonel Ravaska** coule des jours paisibles dans une charmante maisonnette sise dans la verdure, non loin d'Helsinki. […] Le drame a beau être au programme, Arto Paasilinna s'arrange pour faire tourner le vent à l'avantage de son héroïne. Madame **la colonelle Ravaska** n'a pas dit son dernier mot... (Pascale Haubruge, 12 octobre 2001, p. 34.)

Notons qu'aucune des 78 occurrences de *colonel* du corpus ne réfère à une femme.

5.3. *Professions prestigieuses*

Parmi les professions dites de prestige exercée par des femmes, nous avons choisi d'analyser systématiquement certains termes qui relèvent de différents domaines :

5.3.1. *Les sciences et la médecine*

Docteur est une forme particulièrement intéressante. Les 199 occurrences de *docteur* se répartissent comme suit :

	Référent femme	Référent homme spécifique	Référent indéterminé	Emploi générique
docteur	16	160	13	10
doctoresse	13	/	/	/

Seules 16 occurrences réfèrent à des femmes, soit 10 fois moins que celles qui dénotent un référent homme. Si nous analysons de plus près ces occurrences dont les référents sont des femmes, nous obtenons la répartition suivante :

– 2 occurrences sont accompagnées de la marque morphologique du féminin, à travers l'utilisation du déterminant *la* :

> Je n'ai pas un esprit systématique, confie **la docteur** en biologie. (Raphael Duboisdenghien, 17 octobre 2001, p. 23.)

– 8 occurrences sont sans marque morphologique :

> **Docteur** en psychologie à l'UCL, Emmanuelle Zech porte un regard nuancé sur l'émergence de ce phénomène. (Karl Maréchal, 7 septembre 2001, p. 11.)

– 6 occurrences sont accompagnées de la marque morphologique du masculin :

> *Elle dépiste et traite précocement « l'amblyopie fonctionnelle », complète* ***le docteur*** *Marie-Christine Mauroy*, **conseillère** pédiatre et responsable de la banque de données médico-sociales à l'ONE. (Thierry Vanderhaege, 14 décembre 2001, p. 19.)

Dans cet exemple, on remarque que le journaliste féminise la profession de *conseillère pédiatre* pour le même référent dénoté, 4 mots plus tôt, par le masculin *le médecin*. Nous reviendrons dans la section 7 sur les pratiques variationnistes de ce type chez les journalistes.

Par ailleurs, nous avons aussi 13 occurrences de *doctoresse*, forme réputée plus « difficile »[8]. Dans tous ses emplois, *doctoresse* est utilisé dans le sens « docteur en médecine » et non personne porteuse d'un doctorat (en droit, en chimie, etc.). Notons que nous n'avons aucune occurrence de *docteure*.

Sur les 474 occurrences de la forme *médecin*, très proche sémantiquement, seules 5 % réfèrent clairement à des femmes (24 occurrences). Comme pour *docteur*, on trouve soit des marques morphologiques du masculin (*le médecin*), soit des marques morphologique du féminin (*la médecin, une médecin, la jeune médecin*), mais le plus intéressant est sans doute les 9 attestations de la locution *femme médecin* (Voir la section 6).

La forme *chirurgien* est employée 47 fois, pour une seule occurrence de *chirurgienne*. Alors que les occurrences de *chirurgien* se répartissent entre emplois spécifiques (26), génériques (13) ou indéterminés (8), aucun ne réfère à une femme.

En ce qui concerne le couple *chercheur*/*chercheuse*, 17 occurrences réfèrent à des femmes : 12 sous la forme *chercheuse* et 5 (sur les 129) sous *chercheur*. Le plus étonnant est que ces 5 attestations au masculin apparaissent toutes dans le contexte immédiat d'un prénom féminin, comme le montre la concordance suivante. Cette présence d'un prénom féminin aurait pu encourager l'utilisation d'une forme féminisée, ce qui n'est clairement pas le cas :

(1) Par Marie-Cécile Bruwier, docteur en égyptologie, chercheur qualifié au Musée royal de
(2) de sceptiques, enchaîne Marianne Poumay, chercheur attaché au Labset, la cellule en
(3) sant nos cinq sens , confie Françoise Decortis, chercheur du FNRS attachée à l'U.Lg.
(4) le 7 février, t pas moins absent.{S} Pour Anne Wayenberg, chercheur à l'Institut européen
(5) les bonnes questions" {S}Hilde {S}24 ans, chercheur {S}Je sympathise avec Attac sans en

[8] Dans ses recommandations, le Conseil de la langue laisse le choix entre *une docteur* et *une doctoresse*.

Dans les deux premiers exemples, on remarque également les épithètes au masculin : *chercheur qualifié* ou *chercheur attaché*, alors que dans l'exemple (3) l'épithète porte la marque morphologique du féminin : *chercheur attachée*. Cette tournure marque un stade intermédiaire dans lequel le journaliste utilise une séquence mixte, sorte de tâtonnement vers une féminisation pas tout à fait accomplie.

Notons que nous n'avons pas trouvé, en étendant la recherche à la totalité de l'année 2001, la forme québécoise *chercheure(s)*.

5.3.2. *La politique*

La forme *gouverneur* apparaît 171 fois dans notre corpus. Aucun emploi ne réfère à une femme. Par contre, en étendant la recherche à l'ensemble de l'année 2001, nous obtenons les résultats suivants : 9 fois la suite *la gouverneur* (seul le déterminant marque le féminin) et 5 fois la forme féminisée *gouverneure* (le féminin est marqué dans la morphologie du lexème par un *-e* final). Celle-ci est d'autant plus surprenante quand on compare avec l'absence de *chercheure* (cf. ci-dessus).

la gouverneur	9
gouverneure	5

> Le gouverneur de Floride aime sa femme, la **gouverneure** du Massachusetts vient d'accoucher de jumeaux et le ministre de la Justice prie tous les matins. Tout va pour le mieux dans le meilleur des mondes américain ? Si peu... (Nathalie Mattheiem, 22 mai 2001, p. 1.)

Autre terme du domaine politique : *député*. Celui-ci apparaît 353 fois dans le corpus, contre 50 fois pour la forme féminisée *députée(s)*. Seul une forme masculine *député* réfère à une femme dans :

> Nous voulons une garantie supplémentaire, une garantie politique sur la simultanéité, a dit **la député** PS Karin Lalieux. (Philippe de Boeck, 3 octobre 2001, p. 3.)

Si la forme est ici celle du masculin, elle est néanmoins accompagnée du déterminant féminin *la*. Cette tournure semble relativement bizarre, et sa seule apparition dans le corpus nous inciterait à penser qu'il s'agit d'une coquille.

En ce qui concerne le couple du français en Belgique *échevin/échevine*[9], la répartition dans le corpus est la suivante :

	Référent femme	Référent homme spécifique	Référent indéterminé	Emploi générique
échevin	6	372	103	23
échevine	118	/	/	/

À l'exception de 6 cas, les femmes qui possèdent ce titre sont nommées par la forme féminisée (dans 95,2 % des cas). De plus, les attestations référant à des femmes représentent 1/4 des référents spécifiques.

[9] L'échevin est l'adjoint au bourgmestre (maire).

5.3.3. *La justice et le droit*

De la même manière que pour *gouverneur*, aucun des 55 emplois de *huissier* ne dénote un référent de sexe féminin. En 2001 dans *Le Soir*, 2 occurrences de *huissière* apparaissent dans un emploi métalinguistique : en fait, dans un article qui expose les règles de féminisation des noms de profession :

> MODE D'EMPLOI Une camelot, une artisane, **une huissière**, une échevine... [TITRE] […]
>
> Noms terminés au masculin par une consonne. La règle générale d'adjonction du « e » final s'applique. Une avocate, une artisane, une échevine. Avec parfois un redoublement de la consonne (une contractuelle, une pharmacienne) ou l'apparition d'un accent grave (une **huissière**, une préfète). (Nathalie Salengros, 31 mars 2001, p. 46.)

En ce qui concerne *procureur*, la répartition est la suivante :

	Référent femme	Référent homme spécifique	Référent indéterminé	Emploi générique
procureur	22	115	46	33

Sur les 137 emplois spécifiques du corpus, 22 dénotent une femme (soit 16 %). Ces 22 items se répartissent comme suit :

– 7 occurrences sont accompagnées des marques morphologiques du féminin : *la procureur*, *la procureur générale* ;

– 4 occurrences sont sans marque morphologique : le mot est en incise ;

– 11 occurrences sont accompagnées de marques morphologiques du masculin : *le procureur*, *le bureau du procureur*. Parmi ces occurrences, on relève des incohérences de reprise anaphorique, comme dans l'exemple suivant :

> Lotfi Raissi, 27 ans, *était plus particulièrement lié à celui qui pilotait l'avion qui s'est écrasé sur le Pentagone*, a affirmé vendredi **le procureur Arvinda Sambir** devant le tribunal de Bow Street. **Elle** a précisé qu'il pourrait être inculpé *de conspiration en vue de commettre des meurtres*. (AFP, 29 septembre 2001, p. 6.)

Ce genre d'incohérence n'est pas rare, comme le montre l'extrait suivant :

> **Ecrivain, producteur de cinéma, compagne** de Bill Kitteredge, **elle** est aussi la **mère** de deux grands gaillards qui viennent de se lancer dans le cinéma indépendant et présentent leur premier film au Sundance Festival fondé par Robert Redford. (Jean-Marie Wynants, 26 décembre 2001.)

Notons que nous n'avons aucune attestation de *procureure(s)*, même en étendant la recherche à l'ensemble de l'année 2001 du journal.

Toujours dans le domaine juridique, nous obtenons pour la paire *substitut/substitute*, les résultats suivants :

	Référent femme	Référent homme spécifique	Référent indéterminé
substitut	11	38	5
substitute	13	/	/

Ainsi, 11 occurrences parmi celles dont le référent est déterminé (49) dénotent des femmes, soit dans 22,5 % des cas.

Avec les occurrences de *substitute*, on obtient 38,7 % de référents qui sont des femmes parmi les emplois à référent spécifique identifié (24 emplois sur 62). C'est, parmi les termes choisis pour cette enquête, la fonction dans laquelle les femmes sont proportionnellement les plus nombreuses.

5.3.4. *L'armée*

Pour le vocabulaire de l'armée, outre *colonel* (cf. 5.2.) nous avons retenu les termes suivants :

– *lieutenant* : le mot ne dénote aucun référent féminin ; la forme *lieutenante* est absente du corpus (même en l'étendant à l'ensemble de l'année 2001) ;

– *sergent* : sur les 70 occurrences du corpus, une seule réfère à une femme (il s'agit de la poupée Barbie), avec la forme masculine *sergent*.

> Icône immuable, Barbie connaît une certaine forme d'évolution. Elle entre dans la vie comme top, entame les années 80 sur des rollers en chantant du country, puis virage. Elle se met à exercer d'authentiques métiers de mecs. Une année, au hasard, 1992: **sergent** dans le Corps des Marines, docteur, candidate aux présidentielles. Pire, 1995: la voilà maître-nageur, pompier... Physiquement, la blonde subit un relooking tous les 7 ans. (Julie Huon, 19 octobre 2001, p. 14.)

Aucune forme *sergente*, même dans le corpus étendu ;

– *officier* : sur les 131 occurrences de *officier*, une seule réfère à une femme :

> Les premiers élèves ont été les officiers, dit Marianne Van de Keere, **officier** responsable de la formation du commissariat. (Martine Vandemeulebroucke, 15 octobre 2001, p. 7.)

Aucune forme féminisée *officière*, même sur l'ensemble de l'année 2001 ;

– soldat : la forme *soldate*, qui tend à se répandre aujourd'hui dans les médias suite à l'affaire de « Jessica Lynch, la jeune soldate américaine blessée en Irak, capturée puis libérée » (*Le Monde*, 28 décembre 2003) est absente du corpus. Par ailleurs, seule une occurrence de soldat, sur les 83 que compte notre corpus, réfère à une femme, dans l'expression figée « bon soldat » :

> Anne Quevrin, bon soldat convaincu [de la famille royale], a évidemment rectifié le tir […]. (O. Van Vaerenbergh, 27 octobre 2001, p. 31.)

On le voit à travers les termes choisis ici, les femmes sont absentes des métiers de l'armée (pour l'année entière 2001, nous recueillons seulement 3 formes *soldate*).

6. Femme + nom de profession

Nous avons recherché systématiquement dans le corpus les séquences du type : femme(-)nom de profession. Il nous semblait intéressant de recenser les professions concernées par cette structure ainsi que leur morphologie. Les 23 attestations de ce type rencontrées se répartissent comme suit :

– *femme* accompagne 16 occurrences de professions épicènes : *médecin*, mais aussi *commissaire*, *cadre*, *ministre*, *peintre* et *pilote*. Pour ces formes, dont seul le déterminant aurait servi à distinguer le sexe du référent, les auteurs ont ajouté une marque supplémentaire avec le lexème *femme*, semblant par là insister sur le fait que le référent est une femme ;

– *femme* accompagne 4 noms au masculin : *policier*, *soldat* et *écrivain*. Ici, la tournure permet d'éviter les mots *soldate*, *écrivaine* et *policière*, qui ne sont pourtant pas absents de notre corpus (9 *écrivaine* et 10 *policière(s)*) ;

– *femme* accompagne 3 noms de profession employés à la forme du féminin : *femme magistrate* (2 fois) et *femmes potières* (2 fois). La tournure semble ici quelque peu redondante avec l'utilisation des formes féminisées. Pour les cas de *magistrate*, il s'agit pour l'auteur de l'article d'insister sur le fait que la personne est la première femme à exercer cette fonction[10] :

> Le 12 novembre 1948, elle devient officiellement **la première femme magistrate** nommée en Belgique. (Patrice Leprince, 13 novembre 2001.)

Dans tous les cas, ces tournures reflètent l'envie, de la part des journalistes, d'indiquer clairement que le référent est une femme.

7. La variation chez les journalistes

Les trois extraits qui suivent illustrent la variation pratiquée par les journalistes. Si, comme nous l'avons dit, l'emploi des formes féminisées ne relève pas d'une consigne donnée à l'ensemble de la rédaction mais d'un choix personnel du journaliste, il n'est pas rare que celui-ci n'ait pas fixé sa pratique, preuve de l'évolution du phénomène. Sans avoir analysé systématiquement les articles en fonction de leurs auteurs, nous avons pu constater que si, pour certains, l'emploi du féminin pour certaines professions est devenu la norme, d'autres varient d'un article à l'autre, et parfois au sein d'un même article.

Pour le même référent, le journaliste emploie ici les formes *doctoresse*, *femme médecin* et *médecin*, ce dernier n'étant pas féminisé puisque accompagné de marques morphologiques du masculin (*ce* et *il*).

> Anvers Des soupçons de plus en plus lourds **La doctoresse** a peut-être fait une deuxième victime [TITRE]
>
> Une **femme médecin** de 44 ans, interne à l'hôpital Van Enschodt à Willebroek, est soupçonn**ée** d'avoir assassiné un sexagénaire, le 31 juillet 2001. **Elle** est sous mandat d'arrêt depuis le 24 octobre.
>
> **La doctoresse** aurait administré, sans raison apparente, deux injections mortelles (voir « Le Soir » du 12 novembre), mais aurait aussi débranché l'appareil respiratoire. Aucune discussion sur le cas de ce patient n'avait été organisée au sein de l'hôpital et aucune demande d'euthanasie n'avait été formulée.
>
> Mais **ce médecin** a-t-**il**, une semaine auparavant, assassiné de manière identique un homme de 69 ans, de Londerzeel?
>
> C'est ce que pense le fils de ce monsieur qui a introduit une plainte auprès du parquet d'Anvers. Selon lui, l'homme en question n'attendait plus que de quitter l'hôpital quand la **femme médecin** lui a administré deux injections fatales. Le parquet confirme l'enquête sur un éventuel assassinat et cherche à savoir s'il existe des liens entre les deux cas. (Eddy Surmont, 1er décembre 2001, p. 5.)

[10] On retrouve la même manière de procéder dans un article notant systématiquement *gouverneure* et où l'on a *femme gouverneure* dans :

Double bonus : propulsée à la tête de l'Etat, elle en est la première femme — et mère — **gouverneure**. (Nathalie Salengros, 31 mars 2001, p. 46)

Dans l'article suivant, le journaliste alterne *la gouverneur* avec un *madame le gouverneur*.

> Une course attira particulièrement l'attention des foules : la « celebrity race ». Une bonne dizaine de stars se sont prêtées au jeu dont le bourgmestre Jean Demannez (*C'est ça, mon cuistax ?*), le fringant André Lamy (*Comment ça démarre, ce brol ?*), **la gouverneur** du Brabant Véronique Paulus de Châtelet (*Je ne crains rien : j'ai tout organisé pour ma succession*) et le célèbre couple cannois Gaëtan Vigneron (RTBF) – Corinne De Permentier, la bourgmestre de Forest (*Et si ça brûle, il est où, l'extincteur ?*)
>
> […] Quant à **madame le gouverneur** (sixième), elle s'est plantée deux fois dans le virage de « la petite vache ». Commentaire de Demannez, dixième : *Je me devais de laisser passer devant moi* ***la gouverneur****. C'est ma seule véritable opposition à Saint-Josse...*(François Robert, 5 juin 2001, p. 10.)

Dans ce troisième extrait, on peut voir que le journaliste féminise *magistrate* mais que, pour le même référent, il utilise le masculin *substitut*, terme dont le féminin *substitute* n'est pas rare dans notre corpus. Par ailleurs, le journaliste conclut l'article en indiquant *Mlle*, dénotant par là l'état civil de la magistrate[11].

> *Nous considérons cette infraction comme un manque de civisme*, justifie **Françoise Baudru, substitut du procureur du Roi.** […] *Les gens sont assez surpris et expliquent à l'audience qu'ils étaient prêts à payer une transaction*, enchaîne **la magistrate**, qui rappelle que seuls les titulaires d'un document du ministère de la Santé sont autorisés à utiliser ce type de parking.
>
> […] *Il existe évidemment d'autres places pour les gens valides* , conclut **Mle Baudru.** (Nicolas Druez, 29 novembre 2001, p. 18.)

Conclusion

De l'analyse de notre corpus et des termes choisis, nous pouvons dégager quelques constantes. Ainsi, certaines professions semblent toujours exclusivement exercées par des hommes. Parmi l'analyse des noms de professions concernés par cette étude, c'est le cas notamment des métiers de l'armée (*colonel*, *sergent*, *lieutenant*, *soldat*) ou encore de *huissier* ou *chirurgien*. Pour les professions ou fonctions dans lesquelles les femmes sont présentes, on constate que la manière de nommer celles-ci varie grandement d'une fonction à l'autre : ainsi, si pour des termes comme *échevine*, *députée* ou *pharmacienne* les référents féminins sont presque exclusivement dénotés par des noms et déterminants aux marques morphologiques du féminin, un grand nombre d'occurrences dénotant des femmes qui exercent des fonctions prestigieuses sont au masculin. Ainsi par exemple, dans notre corpus, *docteur* et *procureur* sont plus fréquemment accompagnés de déterminants au masculin qu'au féminin, pour des référents féminins. Sans doute la morphologie du mot a-t-elle une incidence sur sa non-féminisation : on trouve par exemple très peu de finales en *–eure* dans notre corpus. On constate par ailleurs que la féminisation semble se faire pour des termes sur lesquels les opposants au décret s'étaient focalisés : *sauteuse* ou *coureuse*, connotés comme sexuels au féminin (plus qu'au masculin ?), se retrouvent dans le corpus. De notre corpus émerge également l'absence de l'usage, aujourd'hui vieilli, qui consiste à donner à une femme le grade féminisé de son mari.

Néanmoins, mis à part quelques cas clairs, la plupart des items que nous avons sélectionnés montrent une grande variation dans les pratiques, non seulement entre journalistes, mais aussi

[11] Le Conseil de la langue recommande l'utilisation généralisée de *madame*.

pour un même journaliste entre différents articles, voire au sein même d'un seul article. Il s'agit sans doute d'autant de façons d'essayer « la bonne formulation ».

Cette variation est la preuve d'un usage encore flottant, d'une évolution en cours. C'est évidemment à l'usager de trancher. Les règles de féminisation proposées par le Conseil de la langue lui permettent de le faire dans un cadre défini, et l'encouragent à reconnaître à la femme, à travers la langue, la place qui est celle qu'elle occupe au XXIe siècle.

Références

Boel E. (1976). Le genre des noms désignant les professions et les situations féminines en français moderne. *Revue Romane* vol. (XI/1) : 16-73.

Brick N. et Wilks Cl. (2002). Les partis politiques et la féminisation des noms de métier. *French Language Studies*, vol. (12) : 43-53.

Fleischman S. (1997). The Battle of feminism and *Bon Usage :*Instituting Nonsexist Usage in French. *The French Review*, vol. (70/6) : 834-844.

Fujimura I. (à paraitre). La féminisation des noms de métiers et de titres dans la presse écrite française de 1988 à 2001. *Mots.*

Houdebine A.-M. (1987). Le français au féminin. *La linguistique*, vol. (23/1) : 13-34.

Khaznadar E. (2000). Masculin et féminin dans la dénomination humaine : linguistique et politique. Aperçu de la pratique québécoise. *Le français moderne* : 141-170.

Larivière L.-L (2001). Typologie des noms communs de personne et féminisation linguistique. *Revue québécoise de linguistique*, vol. (29/2) : 16-31.

Mathieu M.-J. (1999). La France. La féminisation des noms de métiers, titres, grades et fonctions : un bilan encourageant. *La féminisation des noms de métiers, fonctions, grades ou titres. Français& Société*, vol. (10) : 45-63.

Moreau M.-L. (1999). La Communauté française de Belgique. La féminisation des termes de professions en Belgique francophone. *La féminisation des noms de métiers, fonctions, grades ou titres. Français& Société*, vol. (10) : 65-78.

Moreau M.-L. (2001). La féminisation des textes : quels conseils à la politique linguistique ? *Revue PArole*, vol. (20) : 287-313.

Muller Ch. (1994). Du féminisme lexical. *Cahiers de lexicologie*, vol. (65) : 103-109

Planelles Iváñez M. (1996). L'influence de la planification linguistique dans la féminisation des titres en France et au Québec : deux résultats différents en ce qui a trait à l'usage. *Revue québécoise de linguistique*, vol. (24/2) : 71-106.

Wilks Cl. Brick N. (1997). Langue non sexiste et politique éditoriale. *Modern and Contemporary France*, vol. (5/3) : 297-308.

Yaguello M. (1992). *Les mots et les femmes.* Payot.

Yaguello M. (1998). Madame la ministre. *Petits faits de langue.* Seuil : 118-139.

Amélioration de la précision dans un système de question-réponse de domaine fermé

Hai Doan-Nguyen, Leila Kosseim

CLaC Laboratory, Department of Computer Science, Concordia University

Montréal, Québec, Canada, H3G-1M8
{haidoan, kosseim}@cs.concordia.ca

Abstract

This paper presents our experiments in constructing a question-answering system which aims at replying questions on services offered by a large company, here Bell Canada. We concentrate on improving the precision of the information retrieval module of the system. Our approach consists in finding a set of special terms which can effectively characterize the relevance of a retrieved candidate to its corresponding question. Combining this set with the information retrieval module has resulted in very good improvements.

Résumé

Ce papier présente nos expériences dans le développement d'un système de question-réponse qui vise à répondre des questions sur les services offerts par une grande compagnie, ici Bell Canada. Nous nous concentrons sur le problème d'amélioration de la précision du module de recherche d'information du système. Notre approche consiste à découvrir un ensemble de termes spéciaux qui peuvent efficacement caractériser la pertinence d'un candidat à la question correspondante. La combinaison de cet ensemble avec le module de recherche d'information a donné de bons résultats.

Mots-clés : information retrieval, question-answering, recherche d'information, système de question-réponse.

1. Introduction

Ce papier présente nos expériences dans le développement d'un système de question-réponse (QR) (anglais, *question-answering*) qui vise à répondre à des questions d'un client sur les services offerts par une grande compagnie, ici Bell Canada. Cette sorte de système est un exemple de la QR de *domaine fermé* (*closed-domain QA*), qui travaille sur une collection de documents restreints quant au sujet et à la quantité. Le domaine a quelques caractéristiques différentes à la QR de domaine ouvert (*open-domain QA*) qui travaille sur une grande collection de documents de sujets divers, y compris le WWW.

Dans la QR de domaine fermé, les réponses correctes ne peuvent normalement être trouvées que dans très peu de documents. Le système n'a donc pas beaucoup de chance, comme dans le cas de la QR de domaine ouvert, de faire la choix dans un ensemble de candidats abondant de réponses correctes. De plus, si le système vise à répondre aux questions d'un client d'une compagnie, il doit être capable d'accepter des questions complexes, et de formes et de styles librement variés. Il doit aussi retourner des réponses complètes, qui peuvent être très longues et complexes, parce qu'il doit clarifier le contexte du problème posé dans la question, expliquer des options de services, donner des instructions ou des procédures, etc. Ces caractéristiques rendent moins utilisables les techniques développées récemment pour la QR de

domaine ouvert, en particulier, avec les compétitions de TREC (Text REtrieval Conference, par exemple (TREC, 2002)). Ces techniques, qui visent à trouver une réponse courte et précise, se base souvent sur l'hypothèse que les questions sont limitées à des mono-phrases, et peuvent être catégorisées (par exemple en PERSONNE, TEMPS, LIEU, etc.).

La QR de domaine fermé a une longue histoire, qui se commence avec des systèmes travaillant sur une base de données (par exemple, BASEBALL (Green *et al.*, 1961) et LUNAR (Wood, 1973)). Une technique très connue à cause de son originalité était les *semantic grammars* (Brown et Burton, 1975), qui préparent *a priori* des patrons de questions d'une tâche spécifique. Malgré sa simplicité, cette technique ne marche que avec des tâches très petites, et un ensemble limité de questions.

Récemment, la recherche de QR fait attention surtout aux problèmes de la QR de domaine ouvert, particulièrement au problème de trouver des réponses très précises et courtes. Pourtant, on commence à reconnaître la nécessité des réponses longues et complexes. Lin *et al.* (2003) font des expériences prouvant que l'utilisateur préfère une réponse dans contexte, par exemple, un paragraphe qui contient la réponse, sans se soucier de la fiabilité de la source. Buchholz et Daelemans (2001) définissent quelques types de réponses complexes, et proposent une solution où le système présente à l'utilisateur une liste de bons candidats et lui laisse la construction de la réponse. Harabagiu *et al.* (2001) abordent les questions qui demandent une réponse en forme d'une énumération (*listing answer*).

Dans ce qui suit, nous présenterons nos expériences dans la construction d'un système de QR pour Bell Canada, en particulier, les efforts pour améliorer la précision du système.

2. Le corpus et l'ensemble de questions

Le corpus sur lequel nous travaillons contient des documents en anglais qui décrivent les services que Bell Canada offre à la clientèle individuelle et d'entreprise. Ils décrivent les services de téléphone standard, de téléphone sans-fil, d'Internet, etc. La collection de documents a été dérivée d'une version Web de Bell Canada (www.bell.ca). Elle comprend plus de 220 fichiers de texte pur, sans balises à l'intérieur. En général, chaque fichier correspond à une page Web, mais certains fichiers correspondent à plusieurs pages. La plupart des documents sont courts, de 1K à 5K caractères, mais il y a aussi de longs documents (max. 24K caractères). Au total, le corpus comprend environ 560K caractères. Comme le corpus était la dérivation au format texte pur des documents formatés (HTML et PDF), beaucoup d'informations importantes sont perdues, comme les titres, sous-titres, tableaux, énumérations, etc. Dans quelques documents, il reste des informations « bruit », par exemple, des titres des liens de navigation génériques du site Web.

L'ensemble de questions dont nous disposons comprend 120 questions. Il est assuré qu'on peut trouver une réponse dans le corpus pour toute question. La forme et le style des questions varient librement. La plupart des questions sont composées d'une phrase, mais il y en a d'autres qui sont composées de plusieurs. Il y a des "Wh-questions", "Yes-No questions", questions sous forme impérative, etc. Les questions peuvent être pour demander la définition ou un détail d'un service, l'existence d'une sorte de service pour un certain besoin, l'instruction pour une opération concernant un service, etc. Voici quelques exemples de questions :

- *What is Business Internet Dial?*
- *Do I have a customized domain name even with the Occasional Plan of Business Internet Dial?*

- *How can our company separate business mobile calls from personal calls?*
- *Please tell me how I can make an auto reply message when using Bell IP Messaging Webmail.*
- *With the Web Live Voice service, is it possible that a visitor activates a call to our company from our web pages, but then the call is connected over normal phone line?*
- *It seems that the First Rate Plan is only good if most of my calls are in the evenings or weekends. If so, is there another plan for long distance calls anytime during the day?*

L'ensemble de questions a été divisé au hasard en deux parties. La première partie de 80 questions serait utilisée pour le but de développement (*training*). La deuxième de 40 questions serait pour le test final.

3. Recherche d'information

Bien que notre corpus ne soit pas volumineux, il n'est pas non plus si petit pour qu'une recherche directe de réponses des questions dans le corpus soit évidente. De plus, la perte des informations formatées comme les titres, sous-titres, etc., rendrait une telle approche plus difficile. Nous concevons donc notre système suivant la stratégie classique de la QR avec deux étapes : (1) recherche d'information, et (2) sélection de candidats et extraction de réponses.

Pour la première étape, nous utilisons Okapi, un moteur de recherche d'information générique bien connu (www.soi.city.ac.uk/~andym/OKAPI-PACK/, aussi Beaulieu *et al.* (1995)). Après l'indexage du corpus, nous passons les questions de l'ensemble de développement à Okapi. Pour chaque question, nous recevons d'Okapi un ensemble de passages extraits des documents du corpus. Okapi donne aussi un point d'évaluation de la pertinence d'un passage à la question, et le nom du fichier qui le contient – chaque fichier de la collection contribue au maximum un passage pour une question donnée. Les points pour les questions de développement se trouvent entre 1 et 40, et sont précises jusqu'au millième de point, par exemple, 10.741.

Les passages retournés par Okapi sont jugés par un humain d'une manière booléenne : correct ou incorrect. Cette sorte de jugement est recommandée dans le contexte de communications entre une compagnie et sa clientèle. Les conditions et les détails techniques de services de la compagnie devraient être rédigés le plus clairement possible dans la réponse au client. Pourtant, nous acceptons quelques tolérances en jugement. Si une question est ambiguë, par exemple, une question concernant des appels téléphoniques qui ne spécifie pas s'il s'agit du service de téléphone standard ou de téléphone sans-fil, toute réponse correcte correspondant à un cas spécifique serait acceptée. Si un passage est bon mais incomplet en tant que réponse, il sera jugé correct si il contient le thème principal de la réponse souhaitable et que l'humain peut trouver des informations manquantes autour de ce passage dans le fichier contenant.

Le tableau 1 donne des statistiques sur la performance d'Okapi exécuté sur l'ensemble de questions de développement. C(n) est le nombre des passages au rang n qui sont jugés corrects. Q(n) est le nombre des questions dans l'ensemble de développement qui ont au moins un passage jugé correct parmi les premiers n passages. En fait, nous n'avons retenu que 10 premiers passages au maximum pour chaque question, parce que après le rang 10, un bon passage semble très rare.

n	1	2	3	4	5	6	7	8	9	10
C(n)	20	11	5	4	9	3	1	1	4	1
%C(n)	25%	13.8%	6.3%	5%	11.3%	3.8%	1.3%	1.3%	5%	1.3%
Q(n)	20	26	28	32	39	41	42	43	44	45
%Q(n)	25%	32.5%	35%	40%	48.8%	51.3%	52.5%	53.8%	55%	56.3%

Tableau 1. Performance d'Okapi sur l'ensemble des questions de développement (80 questions).

L'estimation ci-dessus a montré que les résultats rendus par Okapi ne sont pas satisfaisants. Bien que le taux de 56.3% de questions ayant une réponse correcte parmi les 10 premiers candidats soit acceptable dans l'état de l'art de QR, les taux pour n de 1 à 5 sont un peu faibles. Malheureusement, ce sont des cas qui correspondent aux buts d'application du système. n=1 signifie qu'une seule réponse serait retournée à l'utilisateur – cela correspond à un système QR totalement automatique. n=2 à 5 correspondent aux scénarios plus réalistes des systèmes semi-automatiques, où un agent de la compagnie choisirait une bonne réponse parmi les n passages retournés par le système, le rédigerait et l'enverrait au client. Nous nous arrêtons à n=5, parce que un nombre plus grand de candidats semble psychologiquement trop pour l'agent humain.

Examinant les passages corrects, nous trouvons qu'ils sont en général assez bons pour envoyer à l'utilisateur comme une réponse compréhensible. Environ 25% des passages corrects contiennent des informations superflues à la question correspondante, tandis que 15% manquent des informations. Pourtant, seulement 2/3 parmi ces derniers (soit 10% du total) se montrent difficiles à être complémentés automatiquement. L'extraction d'une réponse d'un bon passage semble alors moins importante que la tâche d'améliorer la précision du module de recherche d'information du système. Dans ce qui suit, nous nous concentrons sur le problème d'augmenter les Q(n), n=1 à 5, du système.

4. Amélioration de la précision du système

L'idée principale ici est de chercher à pousser le plus possible vers les premiers rangs les passages corrects parmi les meilleurs 10 passages que Okapi retourne pour chaque question. Pour ce faire, il faut trouver une information dans les passages qui a la capacité de caractériser la pertinence d'un passage à la question correspondante. Nous notons que les noms des services spécifiques de Bell Canada, comme *'Business Internet Dial'*, *'Web Live Voice'*, etc., peuvent jouer ce rôle. En fait, ces noms de services apparaissent très régulièrement dans presque tous les documents et questions. De plus, un service est souvent présenté ou mentionné dans très peu de documents, ce qui rend ces termes très discriminatoires. Pour être générique, nous les appellerons les *« termes spéciaux »*. Enfin, dans le corpus, ces termes apparaissent sous forme capitalisée, ce qui nous a permit une extraction automatique facile suivie par un filtrage manuel. Nous avons ainsi obtenu plus de 450 termes spéciaux.[1]

4.1. *Première expérience : le rôle des termes spéciaux*

Pour démontrer que les termes spéciaux peuvent aider à améliorer la performance du système, nous avons conçu une série d'expériences. Premièrement, nous cherchons à savoir si ces termes sont capables d'indiquer la pertinence d'un passage à sa question. Un exemple servira la

[1] Okapi permet un type d'indexage où on peut lui fournir une liste de groupes de mots (deux mots ou plus) comme indices, en plus des indices créés automatiquement à partir de mots seuls. En fait, les résultats présentés dans le tableau 1 correspondent à ce type d'indexage, avec ces termes spéciaux fournis à Okapi. Ces résultats sont bien meilleurs que ceux de l'indexage normal.

présentation de l'idée. Pour la question *'Please explain about Web Hosting Managed Services.'*, le passage 4 est le seul passage qui contient le terme spécial *'Web Hosting Managed'* et, en même temps, le seul passage correct.

Nous concevons donc un système de points basé uniquement sur les termes spéciaux et les mots de catégories grammaticales ouvertes (noms, verbes, adjectifs, adverbes) qui apparaissent en commun dans la question et dans le passage. Nous éviterons de présenter les formules compliquées ici, mais le principe est que plus un passage contient des termes – y compris des termes spéciaux, des groupes nominaux, et des mots de classes verbes, adjectifs, et adverbes – en commun avec la question, plus son point est élevé. Les termes spéciaux contribuent des points plus importants que d'autres termes. Si le passage contient plusieurs fois un terme dans la question, ou deux termes différents dans la question, il recevra une prime (lois de synergie). Nous appellerons ce système de points les *points à termes*. Les passages d'Okapi sont maintenant triés par leurs points à termes, et non pas par les points d'Okapi. Les points à termes donnés par la meilleure formule varient entre 0 et 400.

Nous avons essayé quelques formules différentes, et avec des ensembles de paramètres différents. Les résultats semblent meilleurs qu'Okapi, non remarquablement, mais d'une manière encourageante. Le tableau 2 donne les meilleurs résultats que nous avons obtenus.

n	1	2	3	4	5
Q(n) d'Okapi	20	26	28	32	39
Q(n) du système	24	29	32	34	37
Amélioration	4	3	4	2	-2

Tableau 2. Résultats avec les points à termes.

Une analyse des résultats montre que le système de points à termes a en fait poussé beaucoup de passages corrects vers les premiers rangs, par exemple, il a mis 12 nouveaux passages corrects au rang 1. Pourtant, il a aussi abaissé le rang de beaucoup de passages corrects, par exemple, il en a déplacé 8 hors du rang 1, ce qui a rendu une amélioration totale de 4. Ce mauvais réarrangement sont de deux causes : (1) La longueur des passages qu'Okapi retourne n'est pas homogène : beaucoup de passages corrects sont très courts (par exemple, une ligne), tandis qu'il y a de très longs passages incorrects; et (2) Ce système de points ignore totalement les calculs implicitement effectuées par Okapi. Okapi a peut-être choisi un bon document et extrait un bon passage mais qui ne contient pas beaucoup de termes en commun avec la question correspondante.

4.2. *Deuxième expérience : combinaison de points à termes et points d'Okapi*

Nous nous attendions à ce qu'une combinaison de points à termes et points d'Okapi donne une meilleure performance, c'est pourquoi, nous avons utilisé la formule suivante dans la seconde expérience :

(1) Points du passage = Points à termes + CO * Points d'Okapi

où CO *(coefficient pour Okapi)* est un nombre entier qui prend valeur de 0, 1, 2,... jusqu'à MaxCO. Nous avons choisi MaxCO=60, comme à cet ordre, la partie CO * Points d'Okapi est beaucoup plus grand que la partie Points à termes, et les résultats sont pareils aux ceux d'Okapi originalement.

n	1	2	3	4	5
Q(n) d'Okapi	20	26	28	32	39
Q(n) du système	25	31	36	40	43
Amélioration	5	5	8	8	4
%Amélioration	25%	19.2%	28.6%	25%	10.3%

Tableau 3. Résultats avec les points combinés.

Les résultats présentés dans le tableau 3 ont montré une meilleure amélioration en comparaison avec la première expérience, en particulier pour les n=2 à 5. On remarquera que les résultats pour chaque n correspondent à différentes valeurs de CO.

4.3. *Troisième expérience : rôle de rangs de passages*

Analysant les résultats de la seconde expérience, qui est un peu faible pour n=1, nous trouvons que la partie CO * Points d'Okapi n'a pas bien résolu le problème de mauvais réarrangements dans la première expérience. C'est parce que, dans plusieurs cas, les points d'Okapi pour les passages retournés pour une question ne sont pas très distinctifs, par exemple, ils peuvent être différents l'un de l'autre d'à peine quelques dixièmes de points. Dans ces cas-là, même avec un grand CO, c'est la partie de points à termes qui joue le rôle principal, et le système se comporte de façon identique à celui dans la première expérience. Cette analyse nous conduit à donner au rang des passages dans l'ordre retourné par Okapi un rôle dans le calcul de points. Pour ce faire, nous avons modifié la formule (1) en :

(2) Points du passage[i] = CR[i] * Points à termes + CO * Points d'Okapi

où i est le rang du passage dans l'ordre retourné par Okapi, et CR[i] un *coefficient correspondant au rang*. Le problème maintenant est de trouver de bonnes valeurs pour le vecteur CR.

Plusieurs possibilités sont disponibles. Le cas de distribution normales est CR=(1, 1, 1,..., 1). En notant que la distribution de passages correctes (C(n) dans le tableau 1) se concentre dans les premiers rangs (n=1 à 6), on peut penser à un CR=(1, 1, 1, 1, 1, 1, 0.5, 0.5, 0.5, 0.5). Si on veut pousser plus les passages corrects de rang 2 à 6, on peut utiliser CR=(1, 1.5, 1.5, 1.5, 1.5, 1.5, 0.5, 0.5, 0.5, 0.5). Si on fait attention à la diminution des C(n), on peut proposer CR=(1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1). Si on tient compte de la grandeur relative des C(n) (une perspective probabiliste), on peut penser à CR=(1, 0.55, 0.25, 0.20, 0.45, 0.15, 0.05, 0.05, 0.20, 0.05)[2], etc. En général, on peut construire des CR selon les formes (a), (b), (c) dans la figure 1. La forme (d) donne de mauvais résultats, car elle amplifie les points de passages de bas rangs (de 7 à 10).

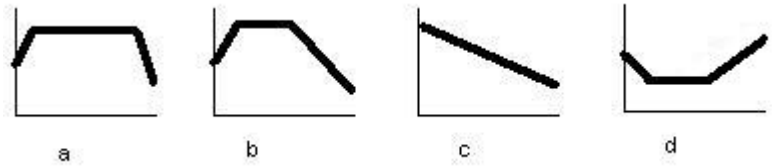


Figure 1. Les distributions possibles pour construire les vecteurs CR.

[2] CR[i]=C(i)/C(1).

Nous avons ainsi construit environ 50 valeurs de CR, et fait fonctionner le système avec ces valeurs et les CO de 0 à 60. Les résultats sont encore meilleurs (tableau 4). Il est intéressant de noter que la meilleur amélioration pour n=1 correspond principalement à la forme (a) de la figure 1 (par exemple, avec CR=(1, 1.2, 1.2,..., 1.2, 0.1), CR=(1, 1.8, 1.8,..., 1.8, 0.1), CR=(1, 3, 3,...,3, 0)). Pour n=2, c'est la forme (b); n=3, la forme (a); n=4, la forme (b); et n=5, toutes les formes (a), (b), (c).

n	1	2	3	4	5
Q(n) d'Okapi	20	26	28	32	39
Q(n) du système	26	33	37	43	43
Amélioration	6	7	9	11	4
%Amélioration	30%	26.9%	32.1%	34.4%	10.3%

Tableau 4. Résultats avec les coefficients de rangs.

4.4. *Quatrième expérience : relation entre la question et le fichier contenant le passage via les termes spéciaux*

Comme la formule (2) n'a pas encore pu résoudre totalement le problème de mauvais réarrangement, dans la quatrième expérience, nous testons la spéculation que si un passage vient d'un fichier qui ne contient aucun terme spécial apparaissant dans la question, ce passage devrait être un « mauvais » candidat, et qu'il faudrait donc le pousser vers les bas rangs, ou même l'éliminer. Cela peut être modélisé en modifiant la formule (2) en :

(3) Points du passage[i] = CT * (CR[i] * Points à termes + CO * Points d'Okapi)

où CT, *coefficient avec termes spéciaux*, sera petit si le passage[i] est « mauvais », et grand si il est « bon ». Nous construisons 20 tels paires de CT pour le test, par exemple (0, 1), (1, 1), (1, 1.5), (1, 2), (1, 2.5), (1, 20), etc. La paire (0, 1) correspond à l'extrémité où tous les mauvais passages sont jetés. Cette fois, nous avons obtenu de très bons résultats (tableau 5). En particulier, les meilleures améliorations pour n=1 et n=2 sont réalisées seulement avec la paire (0, 1). On remarquera que ces meilleures améliorations correspondent aux valeurs non triviales de CO et de CT, ce qui montre que les arguments pour les expériences précédentes restent encore valables. Pour n=3 et n=4, les améliorations ne sont pas les meilleures à la paire (0, 1), mais encore très bonnes à cette paire.

n	1	2	3	4	5
Q(n) d'Okapi	20	26	28	32	39
Q(n) du système	30	38	41	43	44
Amélioration	10	12	13	11	5
%Amélioration	50%	46.2%	46.4%	34.4%	12.8%

Tableau 5. Résultats avec les CT.

4.5. *Test final*

Enfin, nous avons effectué le test final avec les valeurs optimales de CT, CR et CO sur l'ensemble de questions de test (40 questions). Le résultats donnés dans le tableau 6 montrent que le système a fait de très bonnes améliorations avec n=1 et 2, mais moins excellentes avec n=3 à 5. Cela peut s'expliquer par le fait que les passages corrects dans les rangs de 6 à 10 ne

contribuent pas beaucoup à Q(n) : il n'y a que 25-22=3 questions de plus qui reçoivent un passage correcte si on étend n de 5 à 10.

n	1	2	3	4	5	6	7	8	9	10
C(n)	10	6	7	2	3	2	3	2	1	0
Q(n) d'Okapi	10	14	19	20	22	22	24	25	25	25
%Q(n) d'Okapi	25%	35%	47.5%	50%	55%	55%	60%	62.5%	62.5%	62.5%
Q(n) du système	15	19	22	23	23					
Amélioration	5	5	3	3	1					
%Amélioration	50%	36%	16%	15%	5%					

Tableau 6. Résultats du test final.

5. Discussions et conclusions

Dans ce travail, nous avons réalisé des améliorations considérables sur la précision du module de recherche des candidats pour une question. Les termes spéciaux, constitués par les noms de services de Bell Canada, ont joué le rôle principal dans ces améliorations. On peut se demander, pourquoi une telle performance n'avais pas été atteinte par le moteur de recherche d'information (Okapi ici), même si ces termes avaient été entrés au moteur comme termes d'indexage (voir la note en bas de page 1). La raison est peut-être parce que le moteur a traité les termes également comme d'autres termes. En donnant de grands points aux termes spéciaux, ainsi en concevant les coefficients CO, CR, et CT, nous avons rendu le système plus sensible à l'ensemble de termes de travail de l'application.

Le fait que les termes spéciaux ont été extraits facilement dans notre cas, parce qu'ils apparaissent en lettres majuscules dans la collection de documents, ne diminue pas la généralité de l'approche. L'idée principale ici est de trouver une sorte d'information qui peut efficacement caractériser la pertinence d'un candidat à la question correspondante. Cette sorte d'information peut être un ensemble terminologique le plus utilisé dans une application. Il peut être construit manuellement ou (semi-)automatiquement en utilisant de diverses techniques d'extraction.

Pour l'avenir, nous considérerons les approches sémantiques pour notre problème, en remarquant que les noms de services de Bell Canada sont en fait une sorte d'information sémantique. Nous chercherons à savoir si il y a d'autres sortes d'informations sémantiques qui peuvent être utiles, par exemple, le thème de la question, la représentation ontologique des documents, etc.

Remerciements

Ce projet a été financé par les Laboratoires Universitaires Bell (LUB) et le Conseil de Recherche en Sciences Naturelles et Génie du Canada (CRSNG).

Références

Brown J. et Burton R. (1975). Multiple representations of knowledge for tutorial reasoning. In Bobrow et Collins (Eds), *Representation and Understanding*. Academic Press : 311-350.

Buchholz S. et Daelemans W. (2001). Complex Answers: A Case Study using a WWW Question Answering System. *Natural Language Engineering*, Special Issue on Question Answering.

Green W., Chomsky C. et Laugherty K. (1961). BASEBALL: An automatic question answerer. In *Proceedings of the Western Joint Computer Conference* : 219-224.

Harabagiu S., Moldovan D., Pasca M., Surdeanu M., Mihalcea R., Girju R., Rus V., Lactusu F., Morarescu P. et Bunescu R. (2001). Answering Complex, List and Context Questions with LCC's Question-Answering Server. In *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*.

Lin J., Quan D., Sinha V., Bakshi K., Huynh D., Katz B. et Karger D. (2003). The Role of Context in Question Answering Systems. In *Proceedings of the 2003 Conference on Human Factors in Computing Systems (CHI 2003)*, April 2003, Fort Lauderdale, Florida.

Beaulieu M., Gatford M., Huang X., Robertson S.E., Walker S. et Williams P. (1995). Okapi at TREC-3. In Harman D. (Ed.), *Overview of the Third Text REtrieval Conference (TREC-3)*. Gaithersburg, MD: NIST, April 1995.

TREC (2002). *Proceedings of The Eleventh Text Retrieval Conference (TREC 2002)* - NIST Special Publication: SP 500-251, Voorhees E. M. et Buckland Lori P. (Eds).

Woods W.A. (1973). Progress in natural language understanding: An application to lunar geology. In *AFIPS Conference Proceedings*, vol. (42) : 441-450.

Utilisation de Séquences Fréquentes Maximales en Recherche d'Information

Antoine Doucet

Department of Computer Science – PO Box 26 – 00014 University of Helsinki – Finlande
Université de Caen – Département d'Informatique – Campus Côte de Nacre –
14032 Caen Cedex – France

doucet@cs.helsinki.fi, doucet@info.unicaen.fr

Abstract

The growing amount of textual information electronically available has increased the need for high precision retrieval. The use of phrases was long seen as a natural way to improve retrieval performance over the common document models that ignore the sequential aspect of word occurrences in documents, considering them as *bags of words*. However, both statistical and syntactical phrases showed disappointing results for large document collections. In this paper we present a new kind of phrases in the form of *Maximal Frequent Sequences*. Rather mined phrases than statistical phrases, their main strengths are to form a very compact index and to account for the sequentiality and adjacency of meaningful word co-occurrences. They also allow a gap between words. We introduce a method for using these phrases in information retrieval and present our experiments on the INEX collection, a 494 Mb collection of scientific articles. When aggregating the retrieved documents using our phrases with the results of our baseline vector space model system, its average precision for the best 100 documents was improved by 22.8%. The state of the art gives much weaker improvements for similar-sized document collections.

Résumé

La quantité croissante de données textuelles sous forme électronique a augmenté l'importance de la précision des systèmes de recherche d'information. L'utilisation de phrases a toujours été vue comme une technique naturelle pour améliorer la performance des systèmes de recherche. Les techniques classiques sont en effet basées sur des modèles de documents ne tenant pas compte de l'ordre des mots qui les composent. Un document est alors considéré comme un *sac de mots*. Cependant, les phrases statistiques et syntaxiques ont jusqu'ici obtenu des résultats décevants pour des collections de grande taille. Dans cet article, nous présentons les *Séquences Fréquentes Maximales* (SFM), une nouvelle sorte de phrases basées sur des fréquences d'occurrence, mais plutôt issues de techniques de fouille de texte que d'analyse statistique pure. Ces phrases indexent une collection de documents de façon très compacte. En outre, elles tiennent compte de la séquencialité des mots et de leur adjacence, autorisant même un gap entre deux mots formant une même séquence. Nous introduisons une méthode pour exploiter ces phrases en recherche d'information et présentons nos expériences sur la collection INEX, composée de 12,107 articles scientifiques pour une taille totale de 494 Mo. En aggrégeant les résultats obtenus à l'aide des séquences fréquentes maximales à ceux obtenus en utilisant une technique standard, nous améliorons la précision moyenne des 100 premières réponses de 22,8%. L'état de l'art présente des résultats beaucoup plus faible pour des collections de documents d'une taille similaire.

Keywords: statistical phrases, syntactical phrases, information retrieval, text mining.

1. Introduction

Le nombre croissant de documents électroniques rend nécessaires des systèmes de recherche d'information de plus en plus précis. La précision d'un système étant le pourcentage de documents pertinents parmi le nombre total de documents répondus à une requête.

La majorité des systèmes de recherche d'information ne tiennent pas compte de l'ordre des

mots dans un document. Afin d'améliorer leurs performances, il est raisonnable de penser qu'il existe des solutions basées sur cette prise en compte. Zhai *et al.* (1997) évoquent différents types de problèmes causés par l'utilisation de mots simples. Ils constatent notamment que certaines associations de mots ont un sens différent de « l'addition » des sens de ces deux mots (par exemple, l'expression « cordon bleu » ne désigne généralement pas un cordon qui est bleu). Les expressions métaphoriques posent un problème similaire (e.g., « avoir un chat dans la gorge »).

Des travaux sur l'utilisation de phrases en recherche d'information ont lieu depuis plus de 25 ans. Les premiers résultats fûrent très prometteurs. Toutefois, de façon inattendue, la constante augmentation de la taille des collections de documents utilisées a induit une baisse drastique de la qualité des résultats. En 1975, Salton *et al.* (1975) rapportent en effet une amélioration de la précision moyenne sur 10 points de rappel comprise entre 17% et 39%. En 1989, Fagan (1989) réitère les mêmes expériences avec une collection de 10 Mo et obtient des améliorations de 11% à 20%. Cet impact négatif de la taille de la collection fût dernièrement confirmé par Mitra *et al.* (1987), qui pour une collection de 655 Mo n'améliore plus la précision que d'un pourcent ! Turpin et Moffat (1999) revisitent encore ces expériences en 1999 et obtiennent des améliorations de précision incluses entre 4% et 6%.

Une conclusion de ces travaux est que les phrases améliorent les résultats aux bas niveaux de rappel, mais sont globalement inefficaces pour les n premiers documents retournés. D'après Mitra *et al.* (1987), cet apport moindre des phrases pour les meilleures réponses s'expliquent par le fait que leur utilisation promotionne des documents n'évoquant qu'un seul aspect d'une requête. Par exemple, partant d'une requête portant sur les problèmes associés aux fonds de pension, beaucoup des réponses les mieux classés évoquent les fonds de pension, mais aucune difficulté associée. Le problème se rapporte à celui d'une *couverture inadéquate* de la requête.

A notre sens, ceci ne remet pas en cause l'idée qu'ajouter des descripteurs tenant compte de l'ordre des mots doit permettre d'améliorer la performance des systèmes de recherches d'information. Mais ces travaux mettent en évidence le besoin de combiner différemment l'apport des phrases et celui des mots simples (Smeaton et Kelledy, 1998) et surtout le fait que les phrases extraites selon les techniques actuelles ne sont pas satisfaisantes dans l'objectif de représenter les documents d'une collection.

Dans la deuxième section, nous décrivons brièvement le modèle d'espace vectoriel (aussi nommé « sac de mots »), puis présentons les différents types de phrases utilisées dans des travaux liés. En section 3, nous définissons la notion de séquence fréquente maximale et présentons l'algorithme d'extraction correspondant. Nous présentons ensuite la technique de traitement des requêtes correspondante visant à tirer profit des SFM dans le cadre applicatif de la recherche d'information (section 4), avant de présenter notre cadre expérimental et les résultats obtenus (section 5). Nous terminons cet article en tirant les conclusions et en dressant les perspectives prochaines de ce travail (section 6).

2. Utilisation de phrases en recherche d'information

2.1. *Description du modèle d'espace vectoriel*

Représentation des documents. La représentation par le modèle d'espace vectoriel est la plus usitée. Chaque document d'une collection y est représenté par un vecteur à N dimensions, où N correspond au nombre de *caractéristiques* décrivant la collection. Dans la plupart des approches, les caractéristiques sont les mots les plus significatifs.

Un vecteur représentant un document contient le poids de chaque caractéristique dans ce do-

cument. Une valeur fréquemment utilisée pour ce poids est *tf-idf*. *Tf-idf* est une combinaison du nombre d'occurrences du terme dans le document (tf signifie « term frequency ») et de la valeur inverse du nombre de documents dans lesquels il est présent (idf signifie « inverted document frequency »).

La mesure du nombre d'occurrences d'un terme dans la collection (tf) ne permet pas de capturer sa spécificité. Or un terme commun à de nombreux documents est moins utile qu'un terme commun à peu d'entre eux. C'est ce qui motive la combinaison des mesures tf et idf. En bref, tf mesure l'importance d'une terme dans un document, tandis qu'idf mesure sa spécificité dans une collection.

Mesure de similarité. Étant donné une requête et une collection de documents, afin d'obtenir un ensemble de réponses, il est nécessaire de comparer la requête aux documents. Dans le cadre du modèle d'espace vectoriel, on représente la requête par un vecteur appartenant au même espace. Le modèle d'espace vectoriel prend alors tout son sens : il est possible d'utiliser des techniques d'algèbre simple pour calculer des mesures de similarité entre les documents. La mesure la plus fréquente est le cosinus, dont l'atout principal est sa faible complexité. En effet, lorsque les vecteurs sont normalisés, $cosinus(\overrightarrow{d_1}, \overrightarrow{d_2})$ se simplifie en $(d_1 \cdot d_2)$.

2.2. *Utilisation de phrases*

Il existe pour cela différentes possibilités. La méthode usuelle est de considérer les phrases comme des dimensions supplémentaires de l'espace vectoriel, au même titre que les mots simples. Cela pose toutefois quelques problèmes. Le poids donné aux phrases les moins fréquentes est faible. Leur spécificité est pourtant souvent décisive pour déterminer la pertinence d'un document. L'interdépendance entre les différents termes est également problématique. Comment tenir compte du lien entre le poids d'une séquence et celui des deux mots qui la composent ?

Il existe principalement 2 types de phrases, les phrases statistiques, issues d'un simple comptage des co-occurrences de mots seuls, et les phrases syntaxiques.

Phrases statistiques. Pour Mitra *et al.* (1987), une phrase statistique est formée pour chaque paire de 2 mots lemmatisés adjacents et apparaissant dans au moins 25 documents de la collection TREC-1. Les paires sont ensuite triées par ordre lexicographique. Nous relevons ici au moins 2 problèmes. Premièrement, ce classement lexicographique revient à ignorer une information séquentielle précédemment découverte sur une paire de mots : son ordre ! Cela revient à dire que AB=BA. En outre, aucun gap n'est autorisé, il est cependant fréquent de représenter un même concept en ajoutant un mot entre deux autres. Cette définition de phrase ne constate par exemple aucune similarité entre les deux fragments de texte « Université de Basse-Normandie » et « Université de Caen Basse-Normandie ». Ce modèle est très éloigné de la réalité du langage naturel.

Phrases Syntaxiques. La méthode d'extraction de phrase syntaxique de Mitra et al. utilise la nature et la fonction des mots et accepte comme phrases syntaxiques toutes les séquences maximales de mots acceptées par une grammaire découlant d'un ensemble de motifs prédéfinis. Par exemple, une séquence « verbe, nombre cardinal, adjectif, adjectif, nom commun » constituera une phrase de taille 5. Toutes les sous-paires occurrant dans cet ordre seront également générées, avec un gap illimité (par exemple, la paire « verbe, non commun » sera générée). Cette technique permet de très bien représenter le langage naturel. Malheureusement, obtenir nature

et fonction des mots est très coûteux. La taille de l'index est également conséquente, puisque toutes les phrases sont stockées, quel que soit leur nombre d'occurrence. Dans les expériences, Mitra reconnaît en fait ne pas créer d'index pour la collection à priori, mais générer les phrases en réaction à une requête. En pratique, cela signifie un temps d'attente très important pour l'utilisateur. Les résultats sont pourtant similaires à ceux obtenus avec les phrases statistiques.

Nous supposons que cela est dû à plusieurs facteurs. Le premier est certainement l'absence d'un seuil de fréquence minimal pour indexer une phrase. Cela signifie que des phrases très rares ont une influence majeure sur les résultats, alors que leur rareté peut simplement témoigner d'une anomalie. Autoriser un gap illimité pour générer les sous-paires paraît également dangereux : la phrase « I like to eat hot dogs » générera la paire « hot dogs », mais aussi la paire « like dogs », dont le sens sémantique n'a rien à voir avec celui de la phrase initiale.

Séquences Fréquentes Maximales. Nous proposons donc les SFM comme une alternative afin de tenir compte de l'ordre des mots dans la modélisation de documents textuels. Elles présentent l'avantage de n'être extraites que si elles apparaissent avec une fréquence minimale (supérieure à un seuil donné), évitant ainsi l'extraction de phrases non significatives. Un gap entre deux mots est également autorisé au sein même du processus d'extraction, permettant d'appréhender une plus grande variété d'expression.

3. Séquences Fréquentes Maximales

La technique d'extraction des *Séquences Fréquentes Maximales* (SFM) (1999) d'une collection de documents inclut trois étapes principales. L'idée générale respecte les principes de la fouille de données ; sélection et élagage, puis application des techniques centrales du processus de fouille, suivi d'une dernière phase dont le but est de transformer les résultats en connaissances compréhensibles.

3.1. *Définition*

Définition 1. Une séquence $p = a_1...a_k$ est une *sous-séquence* d'une séquence q si tous les items a_i, $1 \leq i \leq k$, occurrent dans q et qu'ils occurrent dans le même ordre que dans p. Si une séquence p est une sous-séquence d'une séquence q, on dit alors que p *occurre* dans q.

Définition 2. Une séquence p est *fréquente* dans une collection de documents D si p est une sous-séquence occurrant dans au moins σ documents de D, où σ est un seuil de fréquence documentaire donné.

Définition 3. Une séquence p est une *(sous-)séquence fréquente maximale* de D s'il n'existe pas de séquence $p' \in D$ telle que p soit une sous-séquence de p', et que p' soit fréquente dans D.

D'après les définitions précédentes, une séquence est dite maximale si et seulement si aucune autre séquence fréquente ne contient cette séquence.

3.2. *Algorithme d'Extraction*

Prétraitement. Cette étape préalable consiste à « nettoyer » les données. Les caractères spéciaux (incluant, par exemple, la ponctuation et les parenthèses) sont effacés. Pour éviter de traiter des items inintéressants, un antidictionnaire est utilisé. Il contient articles, pronoms, conjonctions, adverbes communs, et les formes fréquentes des verbes non informatifs (e.g., « est », « a », « es »). Ces éléments sont ignorés.

Phase initiale : collection des paires fréquentes. Cette phase sert à collecter toutes les paires de mots dont la fréquence documentaire est supérieure à un seuil donné σ. Deux mots forment une paire s'ils occurrent dans le même document, et si la distance qui les séparent est inférieure à un $gap\ g$ donné. Notons également que les paires sont ordonnées, ce qui signifie que les paires (A,B) et (B,A) sont distinctes.

Expansion des paires fréquentes. Pour chaque étape k, $Grams_k$ est le nombre d'ensembles fréquents de longueur k. Ainsi, les paires fréquentes calculées durant la phase initiale composent $Grams_2$. Les SFMs sont trouvées en combinant les séquences fréquentes courtes (de taille k) dans le but de former des séquences plus longues (de taille $k + 1$). Chaque étape inclut de nombreuses phases d'élagage, pour tenter de contrer les risques d'explosion combinatoire.

Au terme du processus, chaque document de la collection est décrit par un ensemble de SFMs.

3.3. *Principaux atouts de la méthode*

La technique permet d'extraire toutes les séquences fréquentes maximales de mots d'une collection de documents. En outre, un *gap* entre les mots est autorisé. Dans une phrase, les mots n'ont pas besoin d'apparaître de façon continue : un paramètre g indique combien d'autres mots deux mots d'une séquence peuvent avoir entre eux. Ce paramètre g est normalement choisi entre 1 et 3.

Par exemple, si g = 2, une phrase « président Bush » est trouvée dans chacun des 2 fragments textuels suivants :

```
...Le président des Etats Unis George Bush...
...Président George W. Bush...
```

Note : Les articles et les prépositions ont été supprimés durant le prétraitement.

L'autorisation d'un gap entre les mots d'une séquence est probablement le plus grand atout de cette méthode, comparée aux autres méthodes existantes pour l'extraction de descripteurs textuels. Cela augmente grandement la qualité des phrases, puisque ce traitement prend en compte la variété du langage naturel.

L'autre spécificité avantageuse des SFMs est la possibilité d'extraire des séquences fréquentes maximales de n'importe quelle taille. Cela permet une description de documents très compacte. Par exemple, en plafonnant la longueur des phrases à 8, une séquence fréquente de mots de longueur 25 nécessiterait plusieurs milliers de phrases pour être représentée (ce qu'un simple calcul combinatoire permet de vérifier aisément).

3.4. *Une technique efficace pour extraire une approximation*

Malheureusement, la présence du gap implique un coup computationnel non négligeable. L'algorithme est exponentiel en le nombre de documents et en leur taille. En pratique, cette exponentialité signifie que pour certains corpus, les SFMs sont calculées en quelques secondes, tandis que pour d'autres collections plus grandes, elles ne peuvent pas être extraites en pratique.

Pour résoudre ce problème, nous avons développé une technique qui permet d'extraire une approximation de l'ensemble des SFMs d'une collection de documents. En divisant la collection de documents en une partition de plusieurs sous-collections, en extrayant l'ensemble des SFM pour chaque collection, et finalement, en joignant chacun de ces ensembles de SFM, nous obtenons une approximation de l'ensemble des SFM de la collection complète.

La validité de cette technique repose sur la conjecture que les SFM extraites sont issues de docu-

ments similaires par nature, et qu'en groupant les documents similaires entre eux, la perte d'information induite par ce partitionnement intermédiaire devrait être minimale. Pour former les sous-collections de documents, nous avons utilisé l'algorithme de clustering *k-means* (Willett, 1988), qui présente l'avantage d'être de complexité linéaire.

En utilisant de petites collections de documents, nous avons pu vérifier la qualité de cette approximation de façon empirique. En ce qui concerne la computabilité, le résultat est net : pour la collection INEX utilisée dans nos expériences, il s'est tout simplement avéré impossible d'extraire les SFM par la méthode directe. En utilisant notre technique de partitionnement intermédiaire, nous avons obtenu des résultats en quelques heures sur un ordinateur familial.

4. Traitement des requêtes

4.1. *Discussion et objectifs*

Etant donné un ensemble de séquences décrivant les documents d'une collection, comment déterminer dans quelle mesure une séquence $p_1 \dots p_n$ décrivant une collection de documents D correspond à une séquence $q_1 \dots q_m$ trouvée dans une requête correspondante ? Et comment conséquemment établir un classement des documents supposés les plus pertinents relativement à cette requête ?

Notre approche consiste à extraire un ensemble de séquences fréquentes décrivant chaque document de la collection. Ces séquences fréquentes sont ensuite comparées aux phrases-clés trouvées dans la requête de l'utilisateur. Chaque document reçoit une *quantité de pertinence* pour chaque séquence qu'il contient correspondant à une phrase de la requête. Ce bonus peut être différent pour chaque phrase.

Il est en effet notamment souhaitable de favoriser les phrases dont l'usage est plus spécifique, en utilisant des coefficients statistiques tenant compte de leur fréquence et de leur spécificité.

Il est également naturel de supposer qu'une plus grande quantité de pertinence découle d'une correspondance plus longue. Si une requête contient la phrase « recherche d'information structurée en XML », il est naturel de privilégier les phrases contenant cette séquence exacte, puis celles en contenant une sous-séquence de taille 3 (par exemple, « recherche d'information structurée »), et enfin celles contenant une sous-séquence de taille 2 (par exemple, « recherche d'information » ou « information structurée »).

Il apparaît également utile de tenir compte du fait que le langage naturel est moins rigide que ne l'est la définition des séquences fréquentes maximales. Il ne serait en effet pas raisonnable d'ignorer la similitude entre une requête ABC et une séquence CBA ou même CAB. On souhaitera cependant prendre en compte le fait que la similarité entre ABC et CAB est plus forte que celle entre ABC et CBA.

Dans l'esprit des séquences fréquentes maximales, nous souhaitons aussi concrétiser numériquement la notion de gap. La phrase AC contient ainsi généralement une forte similitude sémantique avec la phrase ABC (par exemple « information XML structurée » et « information structurée »), quoique dans le cas général, cette similitude avec la phrase ABC est moins forte que celle qui lie la phrase ABC aux paires AB et BC.

Dans la prochaine sous-section, nous présentons la technique que nous proposons et qui tient compte de toutes ces observations.

4.2. *Méthode*

Dans une première étape, nous extrayons les séquences fréquentes descriptives des documents d'une collection suivant la technique décrite en section 3. Pour pouvoir comparer les séquences de mots représentant les documents et les phrases-clés issues d'une requête, nous décomposons les séquences en paires. Ceci permet de comparer des objets de même taille et d'obtenir ainsi des mesures de similarité cohérentes. Chaque paire issue d'une phrase-clé est associée à un score représentant sa *quantité de pertinence*. Cette quantité de pertinence représente « l'importance » de la présence d'une paire de mots dans les documents correspondants. Cette valeur est modifiée par un coefficient d'adjacence qui réduit la quantité de pertinence attribuée par une paire formée de deux mots qui n'apparaissent pas côte à côte dans la phrase-clé dont ils sont issus.

4.2.1. *Définitions*

Soient D une collection de N documents et $A_1 \dots A_n$ une phrase de n mots issus d'une requête contre la collection D. La quantité de pertinence associée à la paire de mots A_iA_j est donnée par :

$$Q_{pertinence}(A_iA_j) = idf(A_iA_j, D) \cdot adj(A_iA_j)$$

où $idf(A_iA_j, D)$ représente la spécificité de A_iA_j dans la collection D :

$$idf(A_iA_j, D) = \log\left(\frac{N}{n}\right)$$

et $adj(A_iA_j)$ est le coefficient d'adjacence visant à pénaliser les paires de mots composées de mots non adjacents dans $A_1 \dots A_n$:

$$adj(A_iA_j) = \begin{cases} 1, & \text{si } A_i \text{ et } A_j \text{ sont adjacents} \\ 0 \leq \alpha_1 \leq 1, & \text{si d}(A_i,A_j) = 1 \\ 0 \leq \alpha_2 \leq \alpha_1 & \text{si d}(A_i,A_j) = 2 \\ \dots & \\ 0 \leq \alpha_{n-2} \leq \alpha_{n-3}, & \text{si d}(A_i,A_j) = n-2 \end{cases}$$

D'évidence on souhaite que $(i \geq j) \Rightarrow (\alpha_j \geq \alpha_i)$, c'est à dire qu'une plus grande distance entre 2 mots implique une moindre pertinence pour la paire correspondante. Dans les expériences, nous nous limiterons à une distance de 1 (i.e., $\forall k > 1 : \alpha_k = 0$).

Notons que la valeur d'adjacence de A_iA_j dans $A_1 \dots A_n$ est aussi nommée le *coefficient modificateur de* A_iA_j.

4.2.2. *Exemple*

En ignorant les distances supérieures à 1, une phrase-clé ABCD d'une requête sera décomposée en 5 couplets (paire, coefficient modificateur) :

(AB, 1), (BC, 1), (CD, 1), (AC, α_1), (BD, α_1)

Comparons cette requête aux documents d_1, d_2, d_3, d_4 et d_5, respectivement décrits par les séquences fréquentes AB, AC, AFB, ABC et ACB (notons ici que le fait que chaque document soit décrit par une seule phrase est un cas particulier qui ne vaut qu'à titre d'exemple). Les quantités de pertinence apportées par la requête ABCD sont indiquées dans le tableau 1.

Document	SFM	Paires correspondantes	Matchs	Quantité de pertinence
d_1	AB	AB	AB	idf(AB)
d_2	ACD	AC CD AD	AC CD	idf(CD) + α_1.idf(AC)
d_3	AFB	AF FB AB	AB	idf(AB)
d_4	ABC	AB BC AC	AB BC AC	idf(AB) + idf(BC) + α_1.idf(AC)
d_5	ACB	AC CB AB	AC AB	idf(AB) + α_1.idf(AC)

Tableau 1. Quantité de pertinence de différentes phrases contre une requête ABCD

On constate que les coefficients modificateurs forment bien un ordre du type souhaité et décrit précédemment. Le seul manque notable est la non prise en compte des paires apparaissant dans l'ordre inverse de celui de la requête (par exemple pour la requête ABCD : BA).

4.3. *Aggrégation des scores de similarité*

Il est évident qu'en pratique, de nombreuses requêtes ne contiennent pas de phrase, et que certaines phrases donneront peu ou pas de résultats. En outre, les documents contenant les mêmes phrases obtiennent tous le même score. Il faut donc les départager afin de pouvoir décider d'un ordre de présentation des résultats à l'utilisateur.

Une idée naturelle serait de re-décomposer les paires en mots simples et de comparer ces mots à ceux de la requête. Cela n'est cependant pas satisfaisant, car les mots les moins fréquents ne peuvent être extraits par l'algorithme d'extraction des séquences fréquentes maximales. Une catégorie de mots potentiellement non extraits, et plus importante encore, est celle des mots qui sont fréquents mais qui ne co-occurrent pas fréquemment avec d'autres.

Pour remplir ce manque, nous extrayons séparemment une valeur de pertinence pour les mots seuls, suivant le modèle d'espace vectoriel présenté dans la sous-section 2.1. Les caractéristiques utilisées sont les mots simples, à l'exception de ceux supposés les moins informatifs, écartés lors d'une phase de prétraitement : petits adverbes, verbes auxiliaires, articles, etc. Il reste alors à combiner les scores de pertinence obtenu par les SFM et ceux obtenus par les mots simples. Pour cela, nous devons d'abord les rendre comparables en les ramenant sur le même intervalle [0,1] grâce à Max_Norm présenté par Lee (1995) :

$$Nouveau\ Score = \frac{Ancien\ Score}{Score\ Maximal}$$

À l'issue de cette première étape, nous combinons les nouveaux scores normalisés en utilisant un facteur d'interpolation linéaire λ, représentant le poids relatif des réponses données par chacune des 2 techniques (Vogt, 1998).

$$Score\ Aggrege = \lambda \cdot score_{Mots\ Simples} + (1 - \lambda) \cdot score_{SFM}$$

Nos expériences nous ont donné de bons résultats en donnant le nombre de mots simples distincts **issus de phrases-clés** de la requête comme poids du score des mots simples, et le nombre de mots simples distincts des phrases de la requête comme poids du score issus des SFM.

5. Expériences et résultats

Nous avons basé nos expériences sur la collection de documents d'INEX (Initiative for the Evaluation of XML retrieval). INEX a vu le jour en 2002 pour répondre à la demande des chercheurs en recherche d'information structurée qui ne bénéficiaient pas jusqu'alors d'un forum

précision@n	Mots simples	SFM	Aggrégé	Amélioration Aggrégé/Mots Simples
précision@10	0,62886	0,53693	0,59933	**-4,7%**
précision@50	0,60918	0,42478	0,58163	**-4,5%**
précision@100	0,05296	0,03467	0,06506	**+22,8%**

Tableau 2. Précision stricte moyenne pour les n premiers documents retournés

d'évaluation spécifique. Une spécificité des documents de la collection INEX est qu'ils ont une structure logique XML fournie. Dans les expériences présentées, nous n'utilisons cependant pas cette structure. La collection INEX se compose d'environ 12,107 articles scientifiques en anglais de l'IEEE, ainsi que d'un ensemble de requêtes et d'assessements compulsés manuellement par les participants. Ces assessements manuels nous permettent d'évaluer numériquement les résultats de notre système.

5.1. *Indexation*

Le premier traitement est d'enlever tous les éléments de ponctuation, les chiffres, les mots de moins de trois lettres, ainsi que ceux appartenant à l'antidictionnaire. Nous extrayons alors les phrases en utilisant un seuil de fréquence de 7, soit la plus petite valeur permettant de calculer l'ensemble des séquences fréquentes maximales en un temps raisonnable.

5.2. *Traitement des requêtes*

Nous n'avons utilisé que les 25 requêtes d'INEX 2002 ne faisant pas cas de la structure XML des documents (CO) et comportant des assessements manuels complets. De ces requêtes, nous n'avons utilisé que les mots et phrases-clés situés dans la balise « Keywords ». Un exemple de contenu d'une telle balise est celui de la requête 47 :

```
<Keywords>
"concurrency control" "semantic transaction management" "appli-
cation"
"performance benefit" "prototype" "simulation" "analysis"
</Keywords>
```

Dans le modèle d'espace vectoriel, nous avons calculé la similarité entre cette requête et les documents de la collection (par exemple, le cosinus des vecteurs correspondants) et classé les documents par score de similarité décroissant.

Pour traiter une requête en utilisant les SFM, nous avons décomposé chaque phrase de la requête en paires comme décrit dans la section 4, en utilisant la valeur d'adjacence arbitraire α_1=0,8 (faire varier ce paramètre n'a qu'une très faible incidence sur les résultats).

Pour obtenir les scores aggrégés, nous calculons λ en fonction du nombre de mots simples total (11 dans l'exemple) et du nombre de mots simples occurrant dans une phrase-clé (7 dans l'exemple). Pour la requête présentée ci-dessus, cela donne $\lambda = \frac{11}{11+7}$.

5.3. *Résultats*

La précision moyenne en considérant les 10, 50 et 100 premiers résultats retournés est donnée en table 2. Cette valeur est obtenue de façon classique à partir d'une courbe de rappel-précision obtenue elle-même suivant la méthode décrite par Raghavan *et al.* (1989) et reprise par Gövert et Kazai comme technique d'évaluation officielle de la première campagne INEX (2003).

Ces résultats confirment le fait que les améliorations apportées par les phrases se situent dans les hauts niveaux de rappel. A un niveau de précision élevé, « l'amélioration » est en réalité une dégradation de la performance.

Il est à noter que le très faible résultat de précision@100 pour les SFM seules s'explique par le fait que beaucoup de requêtes contiennent des phrases qui apparaissent dans moins de 100 documents.

6. Perspectives et conclusions

Nous avons présenté et appliqué un nouveau type de phrases au problème de la recherche d'information documentaire. Nous avons développé et implémenté une technique d'utilisation des séquences fréquentes maximales en recherche d'information. Les expériences menées en utilisant la collection INEX ont donné des résultats encourageants.

Toutefois, l'algorithme d'extraction des SFM est encore lent quand la taille des documents est importante. Cet algorithme a déjà connu de nombreuses améliorations mais il reste certainement perfectible. La technique de partition de la collection en sous-collections permet cependant d'obtenir une approximation en un temps raisonnable.

Il nous faut également expérimenter avec d'autres collections, afin de pouvoir comparer directement nos résultats aux autres. Il est possible que nos résultats soient partiellement dûs à la spécificité de la collection de documents utilisée. Il s'agit en effet d'articles scientifiques, le vocabulaire employé y est donc particulier.

Nos résultats confirment que l'utilisation de phrases améliore les résultats dans les hauts niveaux de rappel. Notre technique est donc certainement plus appropriée à des requêtes d'utilisateurs souhaitant voir la majorité des résultats pertinents. Ce besoin de résultats exhaustifs se trouve par exemple dans le domaine judiciaire, et dans la recherche de brevets.

L'utilisation de phrases est factuelle dans de nombreux langages, ce qui nous rend optimiste quant à de futures expériences avec des corpus multilingues. Le gap doit en outre conférer une robustesse certaine face aux difficultés du multilinguisme.

La découverte des SFM basée sur leur fréquence documentaire reste sans doute un obstacle pour des documents de grande taille. Il serait sans doute opportun de décomposer les articles en sous-documents et de compter le nombre d'occurrences des phrases candidates dans ces sous-documents. Cela donnerait un plus grand nombre de SFM et permettrait d'avoir un poids plus important pour une phrase apparaissant plusieurs fois dans un même document. Différentes granularités sont possible pour définir ces sous-documents et la structure logique de la collection INEX se prête bien à ce type d'expériences.

Références

Ahonen-Myka H. (1999). Finding All Frequent Maximal Sequences in Text. In *16th International Conference on Machine Learning* : 11-17.

Fagan J.L. (1989). The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Science*, vol. (40) : 115-132.

Gövert N. et Kazai G. (2003) Overview of the INitiative for the Evaluation of XML retrieval (INEX) 2002. In *1st INEX Workshop* : 1-17.

Lee J.H. (1995). Combining multiple evidence from different properties of weigthting schemes. *SIGIR* : 180-188.

Mitra M., Buckley C., Singhal A. et Cardie C. (1987). An analysis of statistical and syntactic phrases.

RIAO97, Computer-Assisted Information Searching on the Internet : 200-214.

Raghavan V.V., Bollmann P. et Jung G.S. (1989). A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems*, vol. (7/3) : 205-229.

Salton G., Yang C.S. et Yu C.T. (1975). A Theory of Term Importance in Automatic Text Analysis. *Journal of the American Society for Information Science*, vol. (26) : 33-44.

Smeaton A. F. et Kelledy F. (1998). User-chosen phrases in interactive query formulation for information retrieval. In *20th BCS-IRSG Colloquium*.

Turpin A. et Moffat A. (1999). Statistical Phrases for Vector-Space Information Retrieval. *SIGIR 1999* : 309-310.

Vogt C.C. et Cottrell G.W. (1998). Predicting the performance of linearly combined IR systems. *SIGIR* :190-196.

Willett P. (1988). Recent trends in hierarchic document clustering : a critical review. *Information Processing and Management*, vol. (24/5) : 577-597.

Zhai, Chengxiang, Xiang Tong, Milic Frayling N. et Evans D.A. (1997). Evaluation of Syntactic Phrase Indexing. In *Proceedings of TREC-5* : 347-358.

Spécificités lexicales et acquisition de la terminologie

Patrick Drouin

OLST/ ÉCLECTIK, Université de Montréal, C. P. 6128,
Succursale Centre-ville, Montréal (Québec), H2S 2B6, Canada
patrick.drouin@umontreal.ca

Abstract

In this paper, we present a technique that uses corpus specific vocabulary in order to gain access to terminology. The method exploits a dynamically built heterogeneous corpus made of the merger of a technical and a non technical corpus so as to identify specialized vocabulary in the technical corpus. This highly specialized vocabulary (adjectives and nouns), is validated in order to establish its usefulness for terminology processing.

Résumé

Dans cet article, nous présentons une méthode exploitant le calcul de spécificités dans le cadre d'un processus d'identification de la terminologie. La méthodologie proposée repose sur la constitution dynamique d'un corpus hétérogène (technique/non technique) visant à faire ressortir la trace lexicale laissée dans le corpus technique par la terminologie d'un domaine. Cette trace est identifiée à l'aide du calcul des spécificités. Nous procédons à la validation de la pertinence d'un sous-ensemble des spécificités (nominales et adjectivales hautement spécifiques) afin de vérifier leur utilité dans le cadre du travail du terminologue.

Mots-clés : spécificités, terminologie, acquisition automatique de la terminologie, corpus, analyse de fréquence, langue de spécialité.

1. Introduction

Malgré les progrès récents effectués dans le domaine de l'acquisition automatique des termes (Bourigault *et al.*, 2001 ; Jacquemin, 2001), l'élaboration de dictionnaires spécialisés, bien qu'assistée par ordinateur, demeure une tâche ardue et essentiellement manuelle. Le processus de confection de ces dictionnaires repose encore principalement sur le dépouillement d'une masse de documents spécialisés portant sur un domaine du savoir humain. Le phénomène de plus en plus important de libre circulation des documents en format électronique fait en sorte que le terminologue s'attaque à des corpus de plus en plus volumineux. Les divers documents techniques qui composent ces corpus possèdent bien souvent de nombreuses caractéristiques qui les distinguent des documents non techniques (lexique, style, syntaxe, taille, public cible, etc.). La lecture d'un de ces documents par un non-spécialiste l'amène rapidement à constater la très forte présence de termes techniques.

Nous croyons que la trace lexicale laissée par la terminologie peut être exploitée afin de mettre en lumière les éléments terminologiques contenus dans un corpus de documents techniques. Afin de l'exploiter, nous proposons une méthodologie qui consiste à mettre en opposition le comportement des unités lexicales de corpus de niveaux de spécialisation différents. Pour y parvenir, nous utilisons le calcul des spécificités (Lafon, 1980). Nous ne proposons donc pas un nouvel indice en vue de la description de la spécificité des unités lexicales; nous exploitons plutôt un indice bien connu dans un cadre différent. L'objectif du présent article est de vérifier si l'application du calcul des spécificités à corpus hétérogènes permet

l'obtention de résultats satisfaisants et intéressants pour le travail terminologique.

La section 2 constitue un survol rapide des travaux sur les spécificités, de ceux portant sur des notions apparentées dans le cadre de la terminologie, ainsi que du travail de quelques auteurs qui se sont intéressés à la mise en opposition de corpus dans le but d'identifier la terminologie. La section 3 brosse un tableau de la méthodologie adoptée dans le cadre de nos travaux alors que la section 4 porte sur les résultats obtenus à l'aide du logiciel d'acquisition automatique de la terminologie *TermoStat*.

2. Travaux antérieurs

Dans le cadre de la terminologie traditionnelle et de la terminotique, les études ayant recours à la mise en opposition de corpus sont relativement récentes et peu diffusées. Les approches utilisées dans le cadre de la terminologie se fondent essentiellement sur des analyses de fréquence. Ahmad *et al.* (1994) et Chung (2003) proposent des approches mettant en opposition des corpus dans le but d'identifier le vocabulaire propre à la langue médicale anglaise. Pour sa part, Nelson (2000) s'intéresse à l'identification du vocabulaire du monde des affaires. Il ne s'agit pas à proprement parler d'une étude terminologique et les travaux de cet auteur se rapprochent plus de ceux de Phal (1971) et de Huizong (1986) en ce qu'ils cherchent plutôt à décrire les caractéristiques du vocabulaire d'un ou plusieurs secteurs.

La majorité des études citées précédemment opposent, d'un point de vue de la fréquence d'occurrence, un corpus restreint à un domaine particulier du savoir à un corpus dit « général », composé de documents non liés à un domaine spécifique. Ces études portent sur des documents n'ayant pas l'objet de traitements linguistiques (étiquetage grammatical, lemmatisation, etc.). L'enrichissement des documents peut conduire à des analyses plus fines permettant de distinguer les formes graphiquement identiques, mais possédant des catégories grammaticales différentes. La lemmatisation, pour sa part, permet de regrouper les occurrences d'une forme et d'obtenir une meilleure idée de son importance dans les corpus.

Même si elles exploitent essentiellement la fréquence d'occurrence des formes d'un corpus, les techniques utilisées jusqu'à maintenant sont difficilement comparables. De plus, la variété des seuils de pertinence sélectionnés ne facilite pas la tâche en vue d'une étude comparative. À titre d'exemple, Chung (2003) considère que les formes qui apparaissent 50 fois plus souvent dans son corpus de médecine que dans son corpus général sont dignes d'intérêt. À notre avis, une approche probabiliste, s'éloignant des observations fondées sur les fréquences brutes d'occurrence, pourrait être plus facilement adaptable d'un domaine à un autre et d'une démarche à une autre. Même si l'approche proposée par Lafon (1980) a suscité de l'intérêt dans le cadre du travail en lexicographie (Leselbaum et Labbé 2002 ; Zimina 2002), à notre connaissance, aucune étude n'a envisagé d'avoir recours au calcul des spécificités en terminologie. Nous proposons, dans cet article, une méthode ayant pour but d'évaluer l'utilité des spécificités pour ce type de travail.

3. Méthodologie

3.1. *Description des corpus*

La démarche proposée requiert la constitution de deux corpus, un corpus de référence (CR) et un corpus d'analyse (CA). Afin de tester la stabilité de l'approche, nous reproduirons les tests sur trois corpus d'analyse nommés CA_1, CA_2 et CA_3. Tous les corpus analysés sont en anglais.

La taille totale du corpus de référence est d'environ 7 400 000 occurrences, qui correspondent à approximativement 82 700 formes différentes. Le CR est un corpus non technique composé de 13 746 articles de journaux portant sur des sujets variés tirés du quotidien montréalais The Gazette et publiés entre mars 1989 et mai 1989. Cette diversité de thèmes traités est importante et nécessaire à notre démarche puisqu'elle vient minimiser l'uniformité thématique du CR. On ne peut, bien sûr, s'assurer entièrement qu'un corpus journalistique ne comporte aucune thématique dominante. En effet, les articles qui composent le quotidien sont nécessairement liés à l'actualité et ainsi, à de grandes thématiques sociales. On pourrait aussi envisager de constituer un corpus plus équilibré à partir d'échantillons provenant de documents tirés domaines différents et de documents plus généraux.

Les corpus d'analyse sont de nature technique et correspondent à un seul document. Le CA_1 comporte 11 947 occurrences (1 207 formes), le CA_2 28 583 occurrences (2 066 formes) et le dernier corpus d'analyse (CA_3) est composé de 8 676 occurrences (1 053 formes). Ces corpus présentent donc un éventail de tailles qui rendra possible la validation de la méthodologie sur des ensembles textuels qui possèdent des caractéristiques différentes. Les corpus utilisés pour les expérimentations sont petits, mais nous croyons que leur taille peut être déterminée en fonction des objectifs de travail. Ainsi, dans le cas des documents qui composent les corpus d'analyse, leur taille doit correspondre à un échantillon représentatif traité par les terminologues en situation de travail. Cet objectif, adopté au début de nos travaux, impose une restriction considérable sur le corpus. La taille des corpus d'analyse est donc avant tout dictée par des critères externes.

Bien qu'il soit difficile de classifier catégoriquement un document comme relevant d'un seul domaine de l'activité humaine, nous considérons que les corpus d'analyse traitent du domaine des télécommunications. La nature multidisciplinaire de ce domaine conduit cependant à l'inclusion de concepts venus de différents domaines. Même si celui des télécommunications sert de dénominateur commun aux corpus d'analyse, le corpus CA_1 traite de l'interface de programmation de composantes informatiques. Pour leur part, les corpus CA_2 et CA_3 abordent d'un sujet plus étroit au sein du domaine des télécommunications, celui de la structure physique des réseaux de fibres optiques et de leurs composantes. Le corpus CA_1 s'adresse à des informaticiens qui conçoivent des applications destinées aux composantes présentées dans les documents CA_2 et CA_3. Ces derniers sont rédigés pour des intervenants du domaine des télécommunications ayant une bonne connaissance de la structure physique des réseaux de fibres optiques. Leur public cible est principalement constitué d'architectes de réseaux, d'installateurs, de réparateurs, d'ingénieurs, de testeurs, d'administrateurs de réseaux, etc. Les documents décrivent les possibilités, les caractéristiques, l'entretien, l'utilisation et l'installation des éléments d'un tel réseau.

3.2. *Préparation de corpus*

La première étape de traitement est une segmentation des corpus. L'algorithme de segmentation utilisé est fondé sur celui placé dans le domaine public par Robert MacIntyre de la *University of Pennsylvania*. Le corpus est ensuite étiqueté, sans entraînement préalable, à l'aide de l'étiqueteur conçu par Éric Brill (1992).

Les corpus sont soumis à une étape de lemmatisation heuristique, reposant sur des observations empiriques sur corpus. L'algorithme de lemmatisation consiste à identifier une forme nominale (couple forme/partie du discours ; ex. : *matrices/NNS*), à vérifier si elle comporte un suffixe potentiellement pluriel (ex. : *-ices*), à en retrancher une partie (ex. : -ces), à ajouter un suffixe singulier (ex. : -x), et à rechercher un couple correspondant (ex. : *matrix/NN*) dans la

liste des couples identifiés dans le corpus. Si un couple correspondant est repéré, on considère que la lemmatisation a été effectuée avec succès.

Règle	Suffixe	Retranché	Ajouté	Long. min.	Exemple
1	*-ices*	*-ces*	*-x*	5	*matrices / matrix*
2	*-ives*	*-ves*	*-fe*	5	*knives / knife*
3	*-sses*	*-es*		5	*accesses / access*
4	*-ches*	*-es*		5	*switches / switch*
5	*-eet*	*-eet*	*-oot*	4	*feet / foot*
6	*-ies*	*-ies*	*-y*	4	*possibilities / possibility*
7	*-i*	*-I*	*-us*	4	*stimuli / stimulus*
8	*-s*	*-s*		4	*cars / car*

Tableau 1. Règles de lemmatisation

Nous rejoignons ici la position de Brill (1994) ainsi que de Bourigault et Gonzalez (1994) qui adoptent une approche par apprentissage endogène exploitant le contenu d'un corpus afin de déduire des informations relatives aux autres éléments du corpus. Les règles de lemmatisation présentées dans le tableau 1 nous permettent d'obtenir des résultats satisfaisants. Une analyse d'un échantillon de 1 000 formes nominales prélevées au hasard conduit à une bonne lemmatisation dans 98,7 % des cas.

3.3. *Identification des spécificités*

Nous avons procédé à l'identification des spécificités à l'aide du logiciel d'acquisition automatique des termes TermoStat (Drouin, 2003). La première étape de traitement du corpus par le logiciel en vue de l'extraction des termes est l'identification des spécificités. Ces dernières sont ensuite utilisées à titre de pivots pour la recherche de termes spécifiques au CA. Certaines contraintes sont cependant appliquées au cours du processus d'acquisition et le logiciel ne relève que les spécificités positives nominales et adjectivales. Cette contrainte, bien que très importante, se justifie par la vocation ultime du logiciel et par la nature des termes en langue anglaise qui sont très majoritairement constitués de substantifs et d'adjectifs.

En vue de l'identification des spécificités, TermoStat procède à la constitution dynamique d'un corpus global hétérogène en fusionnant virtuellement le corpus de référence et le corpus d'analyse. Il s'agit ici d'une utilisation inhabituelle du calcul des spécificités qui porte généralement sur un sous-corpus dans le but d'identifier ses spécificités par rapport au corpus d'où il est issu. Nous introduisons donc volontairement un document (le CA) à titre de sous-corpus au sein d'un corpus relativement uniforme (le CR) afin de vérifier dans quelle mesure le comportement des unités lexicales du CA se démarque de ce que l'on observe dans le corpus de référence. Nous avons implémenté au sein du logiciel le calcul de spécificités par approximation normale du calcul hypergéométrique décrit dans Lebart et Salem (1994 : 182).

Du sous-ensemble des spécificités identifiées (nominales et adjectivales), nous restreignons à nouveau le bassin de spécificités et ne retenons que les formes dont la valeur-test est supérieure ou égale à 3,09. Ce seuil minimal nous permet de retenir les formes pour lesquelles nous avons moins d'une chance sur 1 000 d'observer une fréquence égale ou supérieure à celle constatée au sein dans le corpus d'analyse. Il s'agit donc de formes qui sont très forte-

ment représentées au sein du CA et qui, selon nous, devraient être en relation directe avec la trace lexicale laissée par la terminologie dans ce corpus. Aucune contrainte de fréquence minimale n'est imposée aux formes retenues.

4. Résultats

4.1. *Processus de validation*

Notre objectif est de déterminer si les résultats obtenus à l'aide du calcul des spécificités peuvent être utilisés par le terminologue et, si c'est le cas, dans quelle mesure ils peuvent l'être. Afin de valider les données issues de l'acquisition des spécificités, nous avons recours à une banque de terminologie et à des terminologues spécialistes du domaine des télécommunications. La banque de terminologie nous a été fournie par la société *Nortel Networks*, qui a aussi mis à notre disposition les corpus d'analyse. La banque de terminologie comporte essentiellement de la terminologie du domaine des télécommunications. La validation à l'aide de la banque de terminologie consiste en une comparaison *à plat* des listes de spécificités construites à partir des documents qui composent le CA et de la liste extraite de la banque de terminologie. Les formes spécifiques qui sont présentes au sein de la banque de terminologie sont considérées comme pertinentes alors que les autres sont ensuite soumises à une validation humaine afin de juger de leur pertinence. La comparaison purement orthographique effectuée possède des avantages et des inconvénients, mais l'étroitesse du domaine et l'origine commune des documents du CA et de la banque de terminologie nous laissent penser qu'il s'agit d'une approche suffisamment fiable.

Les consignes données aux terminologues pour la validation des formes spécifiques sont simples et ils doivent se limiter à évaluer deux aspects : la pertinence de la forme pour le corpus d'analyse et sa pertinence pour le domaine des télécommunications. Ainsi, si la forme est utilisée dans le domaine des télécommunications ou si elle est représentative du contenu du document, elle est alors considérée comme valide. Ces consignes étendent le champ de validité d'une spécificité à l'ensemble d'un domaine et non seulement au corpus. Certaines unités lexicales pourraient en effet paraître banales au sein d'un corpus, mais elles n'en demeurent pas moins essentielles du point de vue de la terminologie d'un domaine. En effet, le calcul des spécificités ne peut être utilisé que pour déterminer la pertinence d'une forme par rapport à un corpus particulier tiré, dans le cadre de notre démarche, d'un domaine d'activité plus ou moins spécifique.

4.2. *Présentation des résultats*

Le tableau 2 dresse la liste triée en ordre décroissant de valeur-test des 15 premières spécificités pour chacun des trois corpus d'analyse. On remarque que les abréviations sont très nombreuses (*OC*, *OPC*, *ID*, *SDH*, etc.), mais on y trouve aussi des formes pleines et en apparence moins spécifiques (*interface, parameter, amplifier*, etc.) .

Pour sa part, le tableau 3 présente la précision du processus d'acquisition des spécificités, telle qu'elle a été évaluée par l'équipe de terminologues, pour les trois documents qui composent le corpus d'analyse. La bonne performance obtenue doit être interprétée en contexte et en fonction des consignes données lors de l'étape de validation. Ces formes sont, selon les spécialistes, représentatives du corpus ou du domaine des télécommunications.

CA_1		CA_2		CA_3	
Forme	Valeur-test	Forme	Valeur-test	Forme	Valeur-test
interface	192.67	*amplifier*	283.15	*optical (n.)*[i]	257.33
oc	181.02	*optera*	234.11	*opc*	245.26
parameter	173.90	*module*	230.06	*optical (adj.)*	245.00
threshold	167.34	*haul*	211.91	*mor*	226.43
id	158.98	*dwdm*	167.35	*optera*	199.27
ne	150.23	*nm*	164.44	*sdh*	177.78
pm	140.18	*fiber*	163.03	*sonet*	170.81
objectid	137.23	*osc*	162.63	*osc*	169.11
pmbb	137.23	*long*	160.07	*amplifier*	160.26
dn	135.14	*wavelength*	159.82	*input*	154.82
ttp	130.85	*grid*	153.21	*span*	151.46
stm	122.32	*shelf*	150.70	*network*	145.16
invokeid	121.82	*band*	150.39	*haul*	145.15
equipmentid	114.58	*am2*	141.56	*orl*	142.35
attributeid	112.06	*optical*	141.28	*output*	135.32

Tableau 2. Présentation des 15 premières spécificités pour les 3 corpus d'analyse

	CA_1	CA_2	CA_3
Spécificités pertinentes	444	810	273
Spécificités non pertinentes	84	131	101
Précision	84,1 %	86,1 %	73,0 %

Tableau 3. Évaluation de la pertinence des spécificités pour les CA

Tel que nous l'avons mentionné auparavant, les spécificités dont la valeur-test était inférieure à 3,09 n'ont pas fait l'objet d'une validation par les terminologues. On retrouve, dans cette liste, des formes comme *time, rate, process* dans le CA_1, *house, loss, exchange* dans le CA_2 ou encore *point, building, state* et *manager* dans le CA_3. Ces formes sont typiques des documents liés au domaine des télécommunications, mais la mise en opposition des fréquences observées dans les corpus ne leur permet pas de se distinguer suffisamment dans le CA. Dans tous les cas, il s'agit de formes polysémiques ayant un sens non technique (mot) et un sens relevant du domaine des télécommunications (terme). La spécificité de ces formes est donc sémantique et non purement lexicale. Le calcul des spécificités, sans étiquetage sémantique des formes, ne permet malheureusement pas d'identifier cette particularité.

[i] La forme *optical* apparaît deux fois dans cette liste à titre de substantif et d'adjectif. Il s'agit ici d'une erreur d'étiquetage attribuable aux outils informatiques utilisés.

Corpus	Fréquence<=5	Fréquence <=10	Fréquence >10
CA_1	332	430	98
CA_2	664	748	193
CA_3	247	300	74

Tableau 4. Répartition générale des spécificités en fonction de la fréquence

Labbé et Labbé (2001) ont démontré que la fiabilité du calcul des spécificités diminue lorsque la fréquence des événements considérés est basse. Le Tableau 4 donne un aperçu rapide de la répartition des spécificités en fonction de la fréquence dans les corpus traités. On remarque que la majorité des formes recensées ont une fréquence inférieure à 5 (63 % pour CA_1, 71 % pour CA_2 et 66 % pour CA_3). Il est intéressant de noter que notre évaluation par des terminologues de l'intérêt des données, de nature qualitative plutôt que quantitative, met en évidence leur utilité dans le cadre d'une démarche terminologique. L'importance accordée ici à la précision des résultats ne permet pas d'évaluer les performances de l'approche en ce qui concerne le rappel. Il serait intéressant de mesurer, du point de vue du terminologue, l'impact du silence. Cette problématique, beaucoup plus difficile à manipuler puisqu'elle nécessite un dépouillement systématique des corpus d'analyse, devra être abordée dans des études subséquentes.

5. Conclusion

Nous avons présenté une méthode exploitant le calcul de spécificités dans le cadre d'un processus d'identification de la terminologie. Les spécificités sont utilisées comme point de départ pour l'acquisition automatique de la terminologie. Nous avions comme objectif de mesurer la pertinence des spécificités par rapport à un corpus et à un domaine technique. La méthodologie proposée repose sur la constitution dynamique d'un corpus hétérogène visant à faire ressortir la trace lexicale laissée dans un corpus technique par la terminologie. Cette trace lexicale est identifiée à l'aide du calcul des spécificités.

Nous avons procédé à la validation de la pertinence d'un sous-ensemble des spécificités (nominales et adjectivales hautement spécifiques) et il en ressort que, malgré les limites d'une telle approche purement lexicale, ces dernières sont jugées comme utile pour le travail du terminologue. Étant donné les bonnes performances obtenues avec le calcul de spécificités en opposant des corpus de nature différente, nous envisageons de poursuivre nos travaux sur l'acquisition automatique des termes en utilisant les spécificités à titre de pivots pour le recensement des termes.

Références

Ahmad K., Davies A., Fulford H. et Rogers M. (1994). What's in a Term? The semi-automatic Extraction of Terms from Text. In Snell-Hornby, dans Pochhacker M.F. et Kaindl K. (Eds), *Translation Studies. An Interdiscipline.* John Benjamins.

Bourigault D., Jacquemin C. et L'Homme M.C. (Eds) (2001). *Recent Advances in Computational Terminology*. John Benjamins.

Bourigault D. et Gonzalez I. (1994). Acquisition automatique des termes complexes en français et en anglais, approche comparative. In *Proceedings of the Workshop on Compound Nouns: Multilingual Aspects of Nominal Composition* : 29-43.

Brill E. (1994). Some Advances in Transformation-Based Part of Speech Tagging. In *Proceedings of*

the 12th National Conference on Artificial Intelligence (AAAI-94) : 722-727.

Brill E. (1992). A Simple Rule-based Part of Speech Tagger. In *Proceedings of the Third Conference on Applied -Natural Language Processing ANLP-1992* : 152-155.

Chung T. M. (2003). A Corpus Comparison Approach for Terminology Extraction. *Terminology*, vol. (9/2). A paraître.

Drouin P. (2003). Term Extraction Using Non-technical Corpora as a Point of Leverage. *Terminology,* vol. (9/1) : 99-115.

Huizong Y. (1986). A New Technique for Identifying Scientific/Technical Terms and Describing Science Texts. *Literary and Linguistic Computing,* vol. (1/2) : 93-103.

Jacquemin C. (2001). *Spotting and Discovering Terms through Natural Language Processing Techniques*. MIT Press.

Labbé C. et Labbé D. (2001). Que mesure la spécificité du vocabulaire? *Lexicometria*, vol. (3).

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *MOTS*, vol. (1) : 128-165.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Dunod.

Leselbaum J. et Labbé D. (2002). Lexicographie assistée par ordinateur. Signification de « banque » dans le vocabulaire économique In *Actes des JADT 2002* : 447-456.

Nelson M. (2000). *A Corpus-based Study of Business English and Business English Teaching Materials*. Unpublished PhD Thesis, University of Manchester.

Phal A. (1971). *Vocabulaire général d'orientation scientifique*. Crédif.

Zimina M. (2002). Repérages lexicométriques des équivalences à basse fréquence dans les corpus bilingues. *Lexicometria*, n° spécial.

SATO-XML : une plateforme Internet ouverte pour l'analyse de texte assistée par ordinateur

Jules Duchastel[1], François Daoust[2], Dimitri della Faille[3]

[1]Professeur au département de sociologie, UQAM – Montréal – Canada

[2]Informaticien au Centre ATO, UQAM – Montréal – Canada ; doctorant en sciences du langage à l'Université de Franche-Comté – Besançon – France
daoust.francois@uqam.ca

[3]Doctorant au département de sociologie, UQAM – Montréal – Canada
della_faille_de_leverghem.dimitri@courrier.uqam.ca

Abstract

In this contribution, we present a computer-based infrastructure available on the Internet, which allows the manipulation and analysis of text corpora. By the way of an HTML interface the researcher is given access to a personal workspace, a text library, some lexical resources, as well as software applications and procedures for a collaborative work respectful of everyone's data and specific analysis' strategies. The SATO software, available in a client-server mode, allows the categorization of data and the iterative construction of protocols of analysis. XML gives the opportunity to save and exchange data in a standard format. Thus, the described data can be either imported from or exported to other software applications for statistical, linguistic or graphic treatments. The interface available on the Internet includes modes of simplified access to large documented corpora, in particular those of interest for Professor Jules Duchastel's Canada Research Chair in Globalization, Citizenship and Democracy. In this contribution, we are presenting a few exploratory analyses as examples of the possibilities of this computer-based infrastructure.

Résumé

Cet article présente une infrastructure informatique, accessible par le Web, qui permet de manipuler et d'analyser des corpus de textes. Une interface HTML donne au chercheur l'accès à un espace de travail personnel et à des librairies de textes, de ressources lexicales, de programmes et de procédures permettant d'envisager un travail coopératif qui respecte les stratégies d'analyse et les données de chacun. Au niveau des traitements, le logiciel SATO, accessible en mode « client-serveur » permet de catégoriser les données et de construire des protocoles d'analyse de façon itérative. La normalisation XML permet une conservation et un échange des données dans un format standard. Ainsi, les données décrites peuvent être importées ou exportées pour être traitées par divers logiciels statistiques, linguistiques ou graphiques. L'interface Web comprend aussi des modes simplifiés d'accès à de grands corpus documentés, en particulier ceux faisant partie des axes de recherche de la Chaire de recherche du Canada en Mondialisation, citoyenneté et démocratie du professeur Jules Duchastel. Dans cet article, quelques analyses exploratoires illustrent l'utilisation de cette infrastructure logicielle.

Mots-clés : analyse de texte par ordinateur, SATO-XML, interface HTML, corpus sur le Web.

1. Introduction

Le développement d'une infrastructure de recherche au profit de la communauté des chercheurs en analyse de texte vise à rendre accessible sur Internet des *corpus vivants*, c'est-à-dire analysables en ligne en fonction des stratégies spécifiques de chaque chercheur. Nous présentons ici une architecture développée autour du logiciel SATO (Système d'analyse de texte par ordinateur ; Daoust, 1996), mais qui permet également de rassembler divers modules d'analyse statistique, linguistique, etc.

La section deux situe cette architecture dans le contexte du développement du portail ATO-MCD relié à la Chaire de recherche du Canada en Mondialisation, citoyenneté et démocratie du professeur Jules Duchastel (2001). Elle introduit également les principes méthodologiques qui fondent cette architecture. La troisième section est consacrée à la présentation de l'architecture et du modèle SATO. Enfin, dans une quatrième section, nous présentons des exemples d'utilisation de l'infrastructure logicielle accessible par le Web.

2. Contexte et principes méthodologiques

C'est au printemps de 2001 qu'ont débuté les travaux de développement d'une infrastructure de recherche élargie en analyse de texte par ordinateur dans le cadre de la Chaire de recherche du Canada en Mondialisation, citoyenneté et démocratie. Le projet vise à intégrer des acquis développés au cours des années mais qui restent encore trop dispersés (Duchastel, 1993 ; Duchastel et Armony, 1996).

Au niveau méthodologique, cette intégration vise à soutenir une démarche d'analyse de discours. Comme l'écrivait Michel Pêcheux,

> L'analyse de discours ne prétend pas s'instituer en spécialiste de l'interprétation maîtrisant « le » sens des textes, mais seulement construire des procédures exposant le regard-lecteur à des *niveaux opaques à l'action stratégique d'un sujet* [...]. L'enjeu crucial est de *construire des interprétations* sans jamais les neutraliser ni dans le « n'importe quoi » d'un discours sur le discours, ni dans un espace logique stabilisé à prétention universelle (Pêcheux, 1984 ; cité par Maingueneau, 1997).

Comme l'indique Maingueneau,

> Étant donné le statut de l'analyse de discours, on ne peut pas se contenter d'« appliquer » de manière aveugle des protocoles méthodologiques à des corpus. À chaque fois, il faut mener une réflexion spécifique pour construire, de manière interactive, le corpus et son mode d'investigation (Maingueneau, 1997).

L'architecture informatique que nous proposons vise à faciliter cette construction dans un processus itératif et contrôlé dont la trace est explicite.

Au niveau des données, le projet vise l'accueil, la conservation et l'exploitation scientifique de corpus de textes numérisés provenant de la communauté canadienne et internationale des chercheurs en analyse du discours. L'exportation et l'importation des données selon un format XML apparaissent comme une condition pour faciliter la conservation, l'échange et le traitement des corpus et des données lexicales. XML, rappelons-le, est un langage général de balisage des documents électroniques qui permet de publier, conserver, annoter et transformer des textes selon un protocole indépendant des formats propriétaires. Faisant l'objet de concertations (The TEI Consortium, 2001), les protocoles de balisage XML facilitent le transfert des données et des résultats entre logiciels. Signalons que, si le projet vise tout particulièrement les données textuelles en langue française portant sur le discours politique, la plateforme est extensible aux autres domaines de recherche en sciences sociales et en lettres.

Au niveau des traitements informatiques, l'objectif est de fournir un environnement flexible, entièrement accessible via Internet, et permettant au chercheur de déployer ses propres stratégies d'annotation, d'exploration et d'analyse de corpus collectifs ou personnels. Au cœur de la plateforme logicielle, on retrouve le logiciel SATO, augmenté de fonctionnalités permettant l'accueil et l'exportation de données en format XML.

Cette technologie permet d'envisager un véritable travail coopératif jumelant un espace de travail personnel avec des ressources partagées : corpus, bases de données lexicales, documentation et guides méthodologiques. Il sera dès lors envisageable de transformer les collaborations fondées sur le partage de résultats en projets de recherche coopératifs durables voués à

la coexploitation de la base de données et au partage des corpus, lexiques et des savoirs socio-sémantiques. Puisque ce poste de travail électronique utilise une technologie Web standard, il est facilement modifiable et documentable par des tutoriels, manuels, bulles d'aides et guides méthodologiques. Il est également aisé d'implémenter des versions multilingues.

Au niveau matériel, le projet privilégie une approche souple faisant appel à de l'équipement standard rassemblé en îlots de traitement rassemblant plusieurs ordinateurs en réseau. La plateforme logicielle peut donc être déployée dans une variété de configurations allant de l'ordinateur personnel à un réseau élaboré d'ordinateurs se partageant les données et les traitements.

Pour comprendre les motifs à la base de cette stratégie de développement, il faut rappeler les grandes étapes d'évolution des technologies informatiques. On a connu la période des ordinateurs centraux basés, d'une part, sur un traitement centralisé et, d'autre part, sur un accès décentralisé aux données et aux traitements par le biais des terminaux accessibles par modem. Par la suite, on a assisté au triomphe de la micro-informatique qui a démocratisé l'accès au traitement des données sur le poste de travail de l'utilisateur devenu plus puissant que les ordinateurs centraux d'autrefois et à un coût qui dépasse à peine celui des terminaux de jadis. La troisième *révolution* informatique a trait à la généralisation de la réseautique via Internet et l'intégration multimédia et hypertextuelle que permettent le Web et le langage HTML.

HTML est un dialecte issu de la norme SGML. Après avoir connu un développement accéléré et un peu anarchique d'HTML avec la concurrence effrénée dans le développement des navigateurs, le W3C qui arbitre le développement du WEB a décidé d'arrêter l'évolution d'HTML pour promouvoir XML, un langage de balisage issu d'une simplification de SGML et qui intègre la notion d'*extensibilité*. Ce retour à plus de rigueur dans la normalisation des formats des données a pour toile de fond l'impératif de l'échange des données sur Internet dans la perspectives de services Web permettant à des ordinateurs d'échanger des données en vue de les traiter.

Par ailleurs, l'ordinateur personnel est devenu à lui seul un véritable centre de calcul dont l'entretien dépasse souvent les capacités de l'utilisateur, en particulier en ce qui concerne les mises à jour des logiciels et des chaînes de traitement. De plus, dans le domaine de la recherche, nous faisons face à des produits en évolution qui ne disposent pas toujours du même niveau de support que les logiciels commerciaux ou grand public. De là, la nécessité d'aller vers des solutions mixtes qui concentrent des ressources de traitement accessibles par le Web et qui extensionnent le bureau de travail personnel du poste local vers des îlots de traitement distants. De là, aussi, la nécessité du travail coopératif, au-delà du simple échange de publications scientifiques, de telle sorte que l'on puisse échanger des données, en ce qui nous concerne les corpus de textes, les bases de données lexicales, les procédures informatiques et les méthodologies. L'accès à des traitements via le WEB, et la normalisation XML des données à des fins d'échanges entre plateformes informatiques, sont donc des tendances en développement. Outre le projet ATO-MCD, citons, à titre d'exemples d'infrastructure Web dans le domaine de l'analyse des données textuelles à des fins de recherche et d'enseignement, les projets Tapor et Weblex.

Le portail Weblex de l'École normale supérieure Lettres et Sciences humaines de Lyon vise à fournir un accès par Internet à des outils d'analyse textuelle. Encore en développement, le portail permettra un accès aux outils lexicométriques développés depuis des années dans des équipes de recherche dont la tradition remonte au Centre de lexicologie politique de Saint-Cloud. Outre l'accès à des outils d'analyse quantitative des données textuelles aux fonctionnalités apparentées à celles de Lexico (Salem) et Hyperbase (Brunet), le logiciel Weblex entend fournir une édition hypertexte du document et un moteur de recherche très complet

(Heiden, 2002). Le Centre ATO de l'UQAM collabore avec l'équipe de Lyon depuis plusieurs données et la convergence vers des protocoles XML devrait faciliter le transfert des données et des traitements entre les deux groupes.

Au Canada, on retrouve un autre projet de développement d'un portail pour l'analyse textuelle. Il s'agit du Text-Analysis Portal for Research (TAPoR) : « TAPoR permettra l'établissement d'une infrastructure de chercheurs et de ressources informatiques pour l'analyse des textes à travers le pays par la mise sur pied de six centres régionaux afin de former un portail national pour l'analyse des textes » (Rockwell *et al.*, 2002).

Pour sa part, SATO, dans sa version HTML, est offert depuis plusieurs années déjà en accès libre au Centre ATO de l'UQAM à l'adresse http://www.ling.uqam.ca/ato. Tout comme la version DOS qui la précédait, SATO-HTML donne la priorité aux fonctions d'annotation et de catégorisation lexicale et contextuelle ainsi qu'aux stratégies d'analyse personnalisées (scénarios) accompagnées de mécanismes de trace de l'exploration. Comme la plupart des logiciels d'analyse textuelle, on retrouve dans SATO les fonctionnalités classiques de concordance et de fréquences lexicales, mais augmentées de dispositifs de catégorisation. Au niveau des fonctions statistiques, seules les fonctions de base sont directement intégrées au logiciel. En contre-partie, le logiciel permet de produire à loisir des matrices d'occurrences destinées à être traitées par des analyseurs statistiques externes, par exemple des analyses factorielles de correspondance[1].

La section suivante décrit l'architecture du système et ses perspectives de développement futur.

3. Architecture de la plateforme SATO-XML

On pourrait qualifier le logiciel SATO de *tableur textuel*. Le système permet d'accueillir un corpus brut ou déjà annoté ; il permet de l'annoter ou de changer l'annotation déjà présente, de catégoriser le corpus selon des grilles définies par l'analyste et une fois décrit, de l'exploiter de multiples manières. SATO permet de garder une trace complète du processus de description et d'analyse du corpus. Le logiciel offre aussi la possibilité de programmer des dispositifs de *lecture électronique* (Daoust, 2002) et, donc, d'établir des protocoles d'analyse personnalisés et adaptés à chaque type de discours.

SATO, dans ses versions 3 et 4, est un logiciel destiné à supporter une variété de stratégies d'analyse textuelle. Il repose sur une reconfiguration du texte linéaire (chaîne de caractères) sous la forme d'un plan lexique/occurrences. L'axe lexical répertorie l'ensemble des chaînes de caractères constituant les mots, ponctuations, et toutes chaînes de caractères admissibles à un alphabet défini par l'utilisateur. L'axe des occurrences représente l'ordonnancement des unités lexicales suivant l'ordre naturel du texte (de gauche à droite et de bas en haut pour les langues latines).

L'objectif de cette reconfiguration est de faire émerger la dimension lexicale du texte. Il est à noter qu'à part quelques normalisations éditiques mineures, cette reconfiguration est non destructrice, c'est-à-dire qu'elle permet à tout moment de reconstituer le texte original dans sa forme linéaire. Cette reconstitution à la volée permet de produire des éditions sur mesure avec mise en évidence des mots (couleur, soulignement, etc.) selon des critères définis par l'analyste. Il est possible d'exporter ces éditions dans des formats facilitant leur traitement par d'autres logiciels. Aussi, derrière chaque forme lexicale et chaque occurrence, on retrouve un hyperlien donnant accès à diverses fonctions de catégorisation et de parcours.

[1] On trouvera dans le chapitre intitulé « Une stratégie intégrée de recherche en sciences humaines dans le Portail ATO-MCD » un exemple d'intégration de diverses composantes logicielles pour le traitement en chaîne d'un corpus de discours politique.

L'émergence de la dimension lexicale du texte dans le plan lexique/occurrences permettra de distinguer la catégorisation hors contexte, qui appartient au lexique de la langue ou du domaine, de la catégorisation contextuelle, qui appartient davantage à l'énoncé et à la structure discursive. Dans SATO, les systèmes de catégorisation sont appelés *propriétés*. Exception faite de quelques propriétés prédéfinies par le logiciel, l'utilisateur définit lui-même ses propriétés selon les besoins de son analyse.

La catégorisation des formes lexicales ou des occurrences au moyen de ces propriétés peut se faire par manipulation directe à l'écran, précodage sur le texte ou par divers dispositifs algorithmiques : dictionnaires, patrons morphologiques ou filtres sur les propriétés, patrons de cooccurrences positionnelles ou booléennes. Le logiciel permet de constituer ses propres dictionnaires. Des dispositifs d'héritage permettent de définir des propriétés textuelles projetées à partir du lexique ou des propriétés lexicales condensées à partir des occurrences. Le *filtre* est un patron syntaxique permettant de désigner et de rassembler un ensemble de formes lexicales ou d'occurrences par des contraintes sur les caractères de la chaîne ou ses valeurs de propriété.

La définition des contextes pour les concordances, cooccurrences ou segments calculés s'effectue à la volée selon les besoins de l'analyse. On peut aussi définir au besoin des sous-textes et leurs lexiques associés. Le logiciel fournit des dispositifs de comptage permettant de produire diverses matrices d'occurrences dans les segments ainsi constitués. Des mesures statistiques simples permettent de révéler ou de contraster la distribution des fréquences associées aux occurrences spécifiées par un filtre SATO. Les matrices produites par le logiciel peuvent servir de données pour des logiciels d'analyse statistique.

La trace de toutes les manipulations effectuées sur un corpus est enregistrée dans un journal cumulatif daté. On peut, par simple copier-coller des commandes ainsi tracées, constituer des fichiers de commandes appelées *scénarios*. Ces scénarios permettent d'automatiser des fonctions d'analyse et de traitement qui pourront par la suite être appliquées sur divers corpus.

SATO fonctionne en mode client-serveur au moyen d'une interface HTML standard. Le logiciel est accompagné d'un environnement de gestion HTML permettant de définir des comptes d'usagers, d'ouvrir des sessions qui pourront être servies en parallèle. Le système permet de constituer des banques de textes ainsi que des librairies de scénarios et de dictionnaires. L'interface HTML est modifiable à loisir pour créer des applications particulières dans diverses langues. Cette interface permet de jumeler SATO avec d'autres logiciels, des pages HTML et d'utiliser toute la puissance des langages de scriptage comme Perl, PHP, Python, etc.

Les requêtes envoyées par l'utilisateur à partir de son navigateur Web sont d'abord reçues par un programme général, une passerelle, qui gère le dialogue avec une application. Donc, la même passerelle qui dialogue avec SATO peut servir d'interface à tout autre programme qui lit un fichier de commandes et génère un fichier de résultats. Il est donc facile de rassembler autour de SATO une variété de modules informatiques qui seront déployés à la demande de l'utilisateur à partir de son navigateur Web. Ainsi, nous avons déjà mis au point une chaîne de traitement faisant appel au logiciel *Guidexpert* (Plante *et al.*, 2003) pour réaliser une description linguistique et sémantique de corpus. De même, nous prévoyons intégrer des logiciels statistiques et des systèmes de visualisation des résultats commandés par le chercheur dans son espace de travail privé à partir de son navigateur Web.

L'implantation du logiciel dans une architecture Web (SATO-HTML) a permis le développement d'une expertise dans le domaine des interfaces HTML et CGI (*common gateway interface)*. La nouvelle implantation SATO-XML a permis de produire une deuxième version de l'interface qui en augmente l'utilisabilité et qui supporte des interfaces multilingues. Aussi,

toute la partie qui consiste à donner accès au *bureau de travail* de l'usager sur le serveur a été complétée et revue de façon à la distinguer de l'usage du logiciel SATO lui-même. D'autres développements sont à prévoir afin d'exploiter les potentiels de filtrage et de transformation des textes en format XML.

Du point de vue interne au logiciel, la différence la plus importante entre SATO-XML et SATO-HTML sera le passage au jeu de caractères UNICODE, ce qui implique des filtres de conversion permettant de récupérer les données antérieures. Aussi, l'abandon du code hérité de la version DOS sera l'occasion d'augmenter diverses limites de traitement : dimension maximale des corpus, nombre de propriétés, attributs d'affichage et d'hyperliens, etc. Du point de vue de la syntaxe externe des corpus importés et exportés, la nouveauté a trait à l'utilisation de formats XML s'ajoutant au format propriétaire défini avant l'apparition des normes XML et SGML.

On pourrait qualifier la phase actuelle de développement du logiciel de phase de consolidation permettant de passer aux nouvelles normes XML et UNICODE. Ce passage se réalise dans le contexte d'une plateforme de type client-serveur basée sur une technologie Web standard facilitant le traitement coopératif entre logiciels indépendants s'échangeant des fichiers de données dans des formats standardisés. L'étape suivante consistera à ajouter un formalisme et des dispositifs de traitement permettant d'exploiter les relations structurelles tissées par le texte. Les relations les plus immédiates concernent la macrostructure de présentation du texte en sections emboîtées avec titres et renvois. Mais, elles concernent aussi les diverses constructions syntaxiques et stylistiques, les structures argumentaires, rhétoriques, dialogiques, et les divers liens marquant la cohérence textuelle. Ces dispositifs, étudiés par la linguistique textuelle (Adam, 1990), ainsi que la reconnaissance de la *macrostructure sémantique* des textes exigent des dispositifs informatiques de *catégorisation structurelle,* par analogie à la catégorisation simple que nous pratiquons actuellement. L'objectif à plus long terme est donc d'exploiter pleinement les relations entre les segments textuels dans un tracé de *lecture-explicitation* ou dans des analyses lexicales sensibles aux marques de structure.

4. Exemples d'utilisation de la plateforme

Les paragraphes qui suivent illustrent quelques moments d'une analyse réalisée à l'aide de SATO-XML dans son état actuel de développement. Dans notre exemple, nous avons choisi les communiqués de presse en langue anglaise produits par trois groupes de défense des animaux : *World Wildlife Fund*, *Sea Shepherd* et *Greenpeace*. Ces communiqués concernent la levée du moratoire sur la pêche à la morue par l'Union Européenne (en décembre 2002) et l'annonce de la reprise de la pêche à la baleine par l'Islande (en août 2003). Comme ces communiqués ont été émis durant la même période par des groupes différents, mais s'adressant aux mêmes personnes (les membres des groupes, le public en général, les médias ainsi que les organisations mises en cause), ils permettent au chercheur de supposer les groupes assis autour d'une même table installée dans un espace délibératif à l'échelle mondiale, un espace où la production textuelle joue un rôle de premier plan.

Nous avons sélectionné un texte par groupe et par thème (baleines et poissons), soit six textes au total. Il existe pour l'utilisateur deux façons d'envoyer ses textes vers l'espace disque qui lui est réservé sur le serveur : soit à l'aide d'un formulaire disponible dans l'interface du bureau Web de SATO, soit par FTP (*File Transfer Protocol*). L'accès aux textes demeure privé, c'est-à-dire que ces derniers ne sont accessibles qu'à leur propriétaire qui pourra cependant décider de les partager en mode lecture avec d'autres membres de son groupe ou en autoriser le dépôt dans une librairie publique accessible à tous.

Les textes retenus résidant sur le serveur, nous pouvons créer un corpus à l'aide d'un formulaire HTML. Le contenu du corpus sera déterminé par une référence à chacun des six fichiers

contenant les communiqués de presse. SATO en produira alors une représentation sous la forme d'un plan lexique-occurrences. Le logiciel tiendra compte des annotations du chercheur distinguant, par exemple, les diverses parties constitutives des textes : auteurs, titres, sections, etc. L'image 1 (cf. annexe 1) illustre la procédure de soumission d'un corpus.

Cette photo d'écran donne un aperçu de l'interface du bureau sur le serveur. À gauche se trouve le menu. Si on clique sur un item suivi d'un +, on développe les sous-items. Dans cet exemple, on clique sur l'item *Soumission.* La section centrale de l'écran présente la partie supérieure du formulaire de soumission d'un corpus. La section du bas est la bannière d'identification. Pour faire suite à la soumission du formulaire, SATO génère le corpus et passe dans la section analyse du logiciel. Pour les sessions ultérieures, on entrera directement dans la section analyse en choisissant l'item *Corpus personnel* sur le bureau.

Une première manière d'explorer le corpus est d'afficher le lexique des formes lexicales. Dans l'illustration qui suit, nous présentons un lexique ventilé par organisme et par thème. L'image 2 présente l'interface de commande et le formulaire d'affichage du lexique. À gauche, on retrouve le menu de commandes de SATO. En cliquant sur l'item principal *lexique* suivi d'un +, on obtient le formulaire *Afficher* dans la section centrale de l'écran. Le champ *filtre* reçoit alors le patron *$*Freqtot<50>5* qui indique que seuls les items dont la fréquence totale est inférieure à 50 et supérieure à 5 seront retenus. Dans le champ *Tri*, la propriété *Fréqtot* est sélectionnée afin d'ordonner le lexique par la fréquence totale dans le corpus.

L'image 3 (cf annexe 1) présente le résultat de la soumission du formulaire précédent. Outre la colonne indiquant la fréquence totale (Fréqtot), on peut voir une colonne pour chacun des groupes (WWF pour *World Wildlife Fund*, SEA pour *Sea Sheperd* et GRE pour *Greenpeace*), ainsi que la distribution lexicale selon les deux thématiques concernant la pêche à la baleine (BALEINES) et la pêche à la morue (POISSONS). Dans la dernière colonne se trouve la forme lexicale. Dans la partie inférieure de la photo d'écran, on a le journal qui garde la trace de toutes les opérations effectuées durant la session de travail. De plus, la trace cumulative de chaque session est conservée dans le journal associé au corpus. On peut, par simple copier-coller de commandes reproduites dans le journal, construire des scénarios de commandes qui pourront être appliqués à loisir sur le même corpus ou tout autre corpus.

Si on clique sur une forme lexicale, on dévoile dans la fenêtre du bas un menu de catégorisation que nous retrouvons dans l'image 4 de l'annexe 1. La partie droite de la photo d'écran révèle chacune des propriétés associées au mot retenu, ici le nom propre *Lieberman.* On y trouve notamment la propriété *nature_entité* ajoutée en cours d'analyse pour décrire la nature des acteurs sociaux : *technocrates*, *animal*, *protecteurs*, *public*, *autres en faveur des animaux*, *médias*, *scientifiques* et *pêcheurs*. La partie gauche de l'écran de catégorisation contient un menu permettant d'accéder aux contextes courts (KWIC) du mot cliqué, de le catégoriser, de sauvegarder les annotations, etc.

Cette catégorisation sémantique permet de visualiser la fréquence des différents types d'acteurs et leur répartition dans les divers textes. L'image 5 de l'annexe 1 montre un affichage du texte avec mise en évidence des mots correspondants à des acteurs sociaux. À l'écran, les mots sont de couleur différente en fonction de chacune des valeurs de la propriété *nature_entité.* Ces valeurs décrivent les différents acteurs qui s'opposent et s'allient dans le discours des groupes de défense des animaux pour les deux thèmes choisis. Les catégories ont été établies à partir du lexique, mais elles peuvent aussi être désambiguïsées à la lecture du mot en contexte (KWIC). Par exemple, un même acteur peut être considéré, selon le contexte, comme un scientifique ou comme un protecteur.

Un affichage du lexique des catégories d'acteurs (non illustré ici), trié en fonction des fréquences cumulées par groupe et pondéré par la taille du texte, montre que les catégories

d'acteurs qui distinguent le plus les groupes entre eux sont celles d'*autres en faveur des animaux* et de *public* (sur-représenté dans les textes de GRE), de *protecteurs* (sur-représenté dans les textes de WWF) et de *technocrates* (sous-représenté dans les textes de WWF).

Il apparaît, après l'affichage du texte catégorisé et coloré (non illustré ici) en fonction des différentes valeurs de la propriété *nature_entité*, que le discours de *Greenpeace* met en scène le plus grand nombre d'acteurs, correspondant à l'ensemble des valeurs de la propriété et répartis également dans le sous-corpus. Quant au *World Wildlife Fund*, reconnu comme le moins radical des trois groupes, il insiste dans son texte sur la morue, sur les *protecteurs* des animaux, la présence des autres acteurs n'étant que suggérée alors que le texte concernant les baleines mentionne le *public* (*community, people, consumers*). *Sea Shepherd* n'évoque, dans son texte sur les baleines, que les *pêcheurs* (*fleet*, *whalers*) confrontés à l'action directe du groupe. Le texte sur la morue est moins menaçant et moins direct et le nombre d'acteurs mentionnés s'accroît. L'opposition mise en évidence par le groupe se trouve cette fois entre le *public* et les *gouvernements*. Ces données confirment le radicalisme reconnu du groupe qui n'hésite pas à saborder les baleiniers.

Divers lexiques d'occurrences ou de cooccurrences peuvent être générés en fonction des critères de partition du corpus. Plusieurs tableaux ont été produits par des analyseurs statistiques simples appliqués sur le corpus. Les limites de cet article ne permettent pas de les reproduire ici. Il s'agissait plutôt d'illustrer quelques moments d'une analyse et de mettre en lumière les possibilités d'une plateforme Internet ouverte. On pourra, par ailleurs, consulter une démonstration en ligne sur le site web de la chaire MCD et du Centre ATO de l'UQAM.

Références

Adam J.-M. (1990). *Éléments de linguistique textuelle, Théorie et pratique de l'analyse textuelle*. Mardaga.

Brunet Ét. *Logiciel HYPERBASE (version 2.3*), http://ancilla.unice.fr/~brunet/pub/hyperbase.html site visité le 7 janvier 2004.

Daoust F. (1996). *SATO 4, Manuel de référence*. Centre d'analyse de texte par ordinateur, UQAM.

Daoust F. (2002). L'analyse de texte assistée par ordinateur, lunette de lecture des textes électroniques. *Communication présentée au colloque Publications et lectures numériques : problématiques et enjeux, 70ième congrès de l'ACFAS*, EBSI, Montréal. http://www.ebsi.umontreal.ca/rech/acfas2002/daoust.pdf

Duchastel J. (1993). Discours et informatique : des objets sociologiques ? *Sociologie et sociétés*, vol. (25/2) : 157-170.

Duchastel J. (2001). *Présentation du projet ATO-MCD.* http://ato.chaire-mcd.ca/presentation/

Duchastel J. et Armony V. (1996). Textual Analysis in Canada: An Interdisciplinary Approach to Qualitative Data. *Current Sociology*, vol. (44/3) : 259-278.

Heiden S. (2002). *Weblex, Manuel Utilisateur, version 4.1 intermédiaire*, http://lexico.ens-lsh.fr/doc/weblex.pdf https://weblex.ens-lsh.fr/ sites visités le 7 janvier 2004.

Maingueneau D. (1997). *L'Analyse du Discours*. Hachette.

Pêcheux M. (1994). Sur les contextes épistémologiques de l'analyse du discours. *Mots*, vol. (9). Presses de la Fondation nationale des sciences politiques.

Plante P. *et al.* (2003). Guidexpert ATO. http://fable.ato.uqam.ca/guidexpert/guidexpert-ato-wp.htm

Rastier F. (1989). *Sens et textualité*. Hachette.

Rockwell *et al.* (2002). *TAPoR: Text-Analysis POrtal for Research*. http://huco.ualberta.ca/Tapor/ site visité le 7 janvier 2004.

Salem A., Lamalle C., Martinez W. et Fleury S. *Lexico 3*. http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/ site visité le 7 janvier 2004.

The TEI Consortium (2001). *Text Encoding Initiative, The XML Version of the TEI Guidelines*. In Sperberg-McQueen C.M. et Burnard L (Eds).

Annexe 1 : Photos d'écran

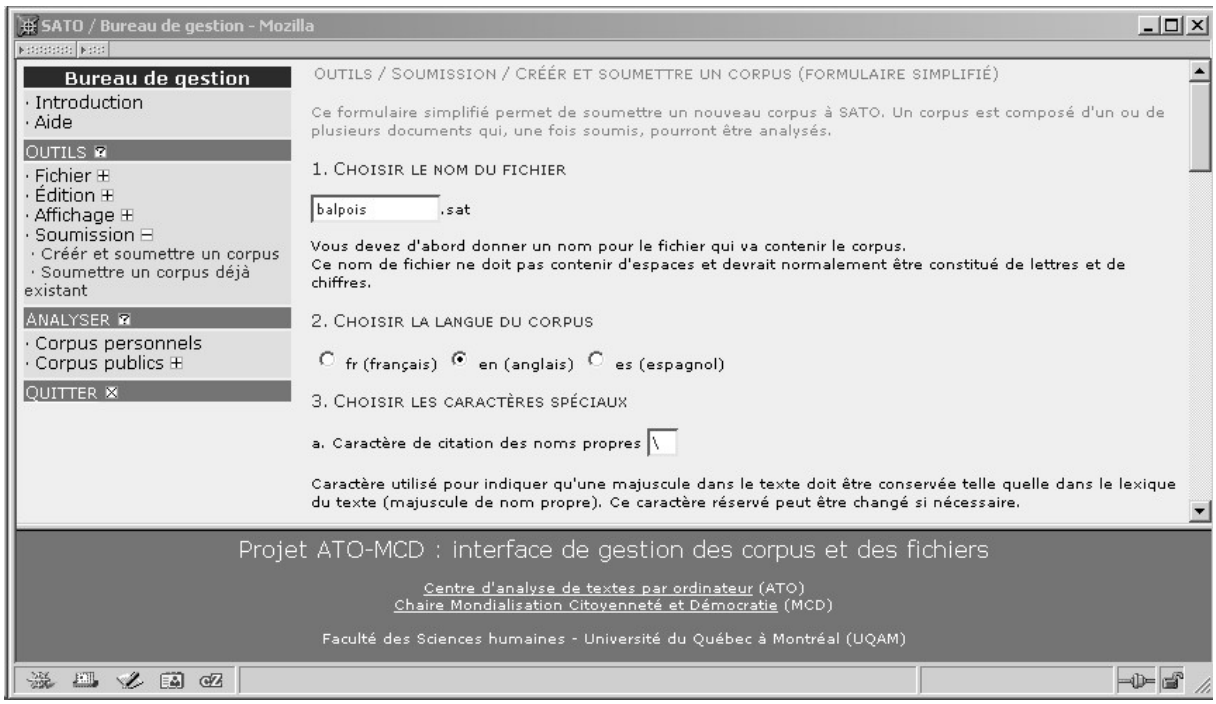


Image 1. Soumission d'un corpus

Image 2. Formulaire d'affichage du lexique

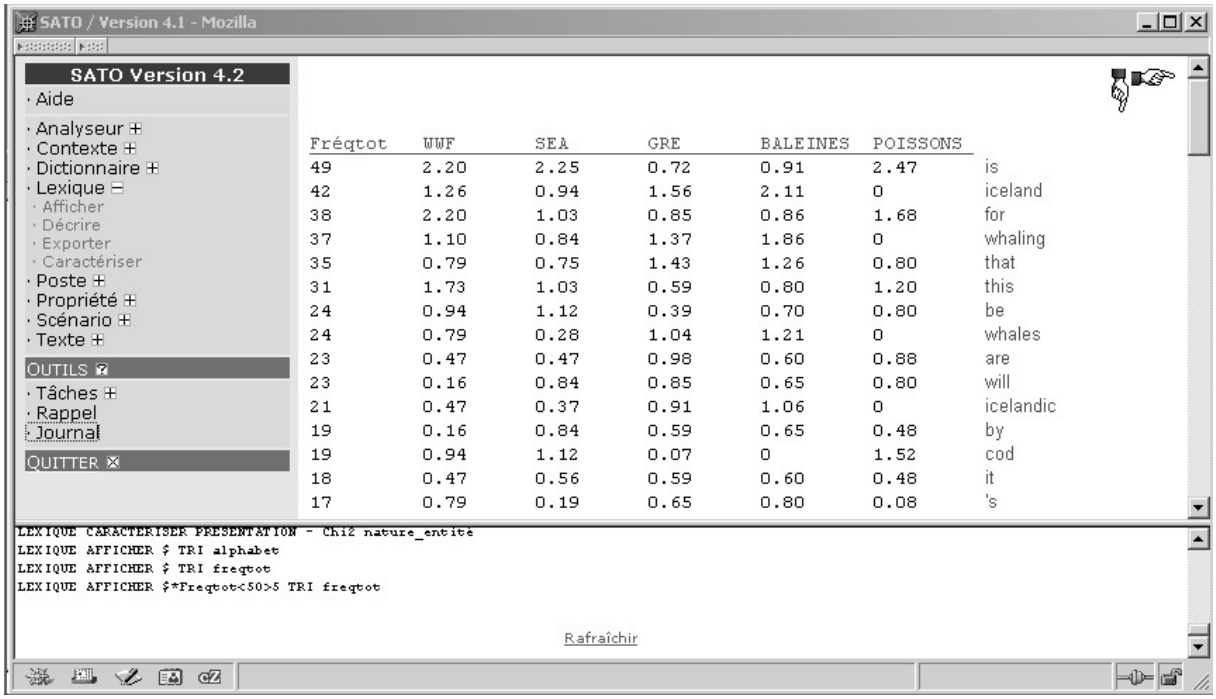


Image 3. Affichage du lexique et du journal

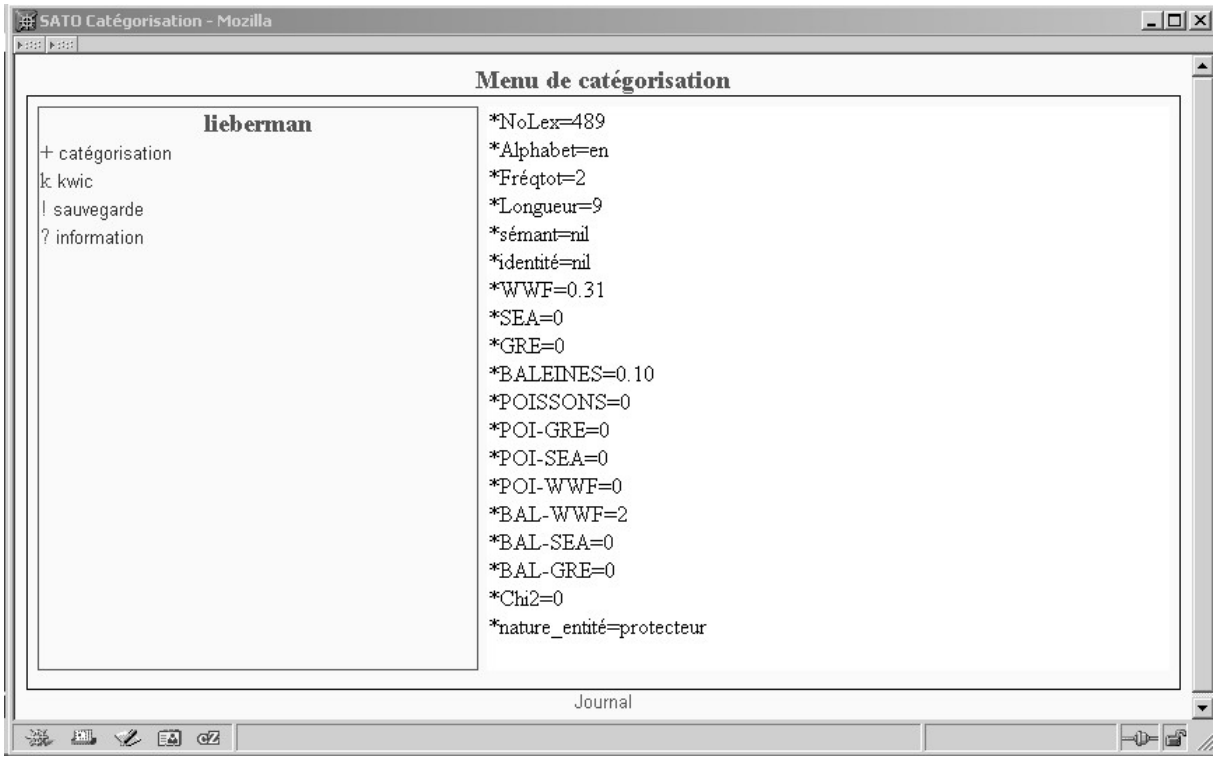


Image 4. Menu de catégorisation avec l'information

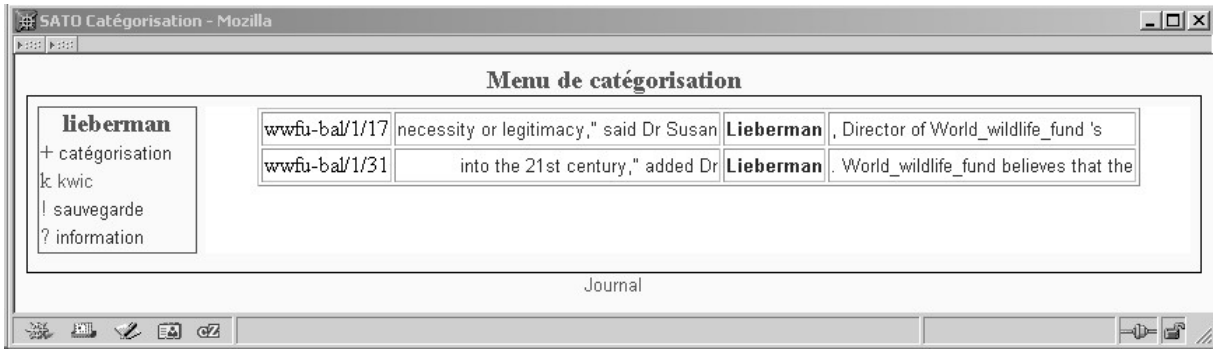


Image 5. Menu de catégorisation avec le KWIC

Image 6. Affichage du texte avec mise en couleur des acteurs sociaux

Une stratégie intégrée de recherche en sciences humaines dans le Portail ATO-MCD

Jules Duchastel[1], Francis J. Lacoste[2], François Pizarro Noël [3]

[1]Titulaire de la Chaire de recherche du Canada en Mondialisation, Citoyenneté et Démocratie, Professeur titulaire, département de sociologie, Université du Québec à Montréal
[2]Chercheur-programmeur à la Chaire MCD – francis@contre.com
[3] Responsable méthodologique à la Chaire MCD – francois@contre.com

Abstract

This paper presents the ATO-MCD Portal, a Web-based cooperative environment for research projects in content and discourse analysis. The paper is a demonstration of this environment. As such, it showcases the functionalities of the Portal, which are designed to work on corpus and develop analytical grids. It also showcases statistical tools for the analysis of frequency tables. The paper explains how those functionalities and tools are embedded in a Project Book that facilitates coordination between researchers and helps to document the research process. We will present its functionalities designed for the treatment of the corpus and the construction of analytical frameworks. We'll then present the statistical tools meant for the analysis of huge frequencies tabs. The presentation of the paper will be a presentation of this research environment.

Résumé

Cet article présente le Portail ATO-MCD, un environnement Web coopératif pour des projets de recherche en analyse de contenu. Nous présentons les fonctionnalités du Portail pour le traitement des corpus, la construction de grilles d'analyse et les outils statistiques pour l'analyse des tableaux de fréquence. Nous montrons comment ces fonctionnalités sont intégrées autour d'un Cahier de projet qui facilite la coordination entre les chercheurs et favorise la documentation du processus de recherche. La présentation consistera en une démonstration de l'environnement.

Mots-clés : analyse de texte par ordinateur, analyse de contenu, analyse factorielle des correspondances.

1. Introduction

Le Portail ATO-MCD est un environnement coopératif de recherche réalisé par la Chaire de recherche du Canada en mondialisation, citoyenneté et démocratie, le GRADiP (Groupe de recherche en analyse du discours politique) et le centre ATO (Centre d'analyse de texte par ordinateur) dans le cadre du projet ATO-MCD (http://ato.chaire-mcd.ca/). Le projet comporte deux volets. Le premier volet documentaire met à la disposition du grand public et de la communauté de chercheurs un grand nombre de corpus reliés à l'univers du discours politique. La banque de données initiale est composée des corpus constitués au cours des années dans les recherches du GRADIP. En effet, les corpus DNL (Discours Néo-Libéral : documents provenant des gouvernements élus, des syndicats, des conseils patronaux, des Églises et des organismes gouvernementaux), EDM (Espace Délibératif Mondial : comportant des textes des grandes organisations internationales) et CCC (Corpus Constitutionnel Canadien : discours des Premiers ministres fédéraux et provinciaux lors des conférences constitutionnelles, 1940-1992) sont dores et déjà disponibles sur le Portail. À la différence d'une banque tex-

tuelle classique, non seulement le texte des documents constituant les corpus sont accessibles, mais les grilles d'analyse et les notes des chercheurs sont parties intégrantes de ces banques.

Outre ce volet « accessibilité aux fruits de la recherche », le projet ATO-MCD possède aussi un volet « instruments de recherche ». En ce sens, il se veut un environnement coopératif qui intègre à l'intérieur d'une même interface un ensemble d'outils méthodologiques pour l'ATO. Les outils méthodologiques intégrés dans le projet correspondent à la perspective d'« analyse de contenu » privilégiée dans les recherches de la Chaire[1].

Nous insisterons ici sur la dimension « environnement coopératif de recherche ». Nous aborderons tout d'abord les principes généraux qui ont orienté la conception de cet environnement. Afin d'accroître l'aspect coopératif, nous avons renforcé les fonctionnalités favorisant une bonne documentation du processus de recherche. Les fonctionnalités reliées à la gestion des documents et des corpus seront ensuite présentées avant d'aborder les différents outils pour le traitement des données et la production d'analyses. Nous terminerons en esquissant quelques orientations futures pour le développement du Portail.

Pour illustrer notre démarche, nous présenterons l'utilisation que nous avons faite du Portail pour mener à bien un projet de recherche portant sur les discours d'ouverture des sessions législatives par les différents Premiers ministres du Québec entre 1960 et 2003. Le but de cette recherche était d'émettre une « opinion informée » dans la polémique sur la reprise de la « Révolution tranquille » telle qu'évoquée par le nouveau Premier ministre libéral, M. Charest, dans son discours d'ouverture de la session législative à l'automne 2003.

Le projet « Ouverture » nous aura permis de développer une stratégie intégrée de recherche dans le cadre du Portail ATO-MCD en fonction de nos besoins réels de chercheurs. Cette étude d'ampleur restreinte nous a permis ainsi de couvrir chacune des étapes du processus de recherche. Cela nous a conduit à mieux évaluer l'efficacité et la pertinence des outils que nous souhaitons mettre en œuvre sur le Portail.

Le projet Ouverture porte sur les discours d'ouverture tenus à l'Assemblée Nationale du Québec, de la première législature de Jean Lesage en 1960 jusqu'au discours de Jean Charest en 2003. Le corpus DNL comprenait déjà les discours d'ouverture prononcés entre 70 et 84. Il suffisait de compléter le corpus en y incorporant les discours de la décennie 1960 et ceux de la période 1985-2003.

Le corpus additionnel provient de deux sources documentaires :

1. Avant 1964, les *Journaux de l'Assemblée Législative de la Province de Québec* qui contiennent une reconstitution des débats parlementaires. Le discours d'ouverture, prononcé au nom du gouvernement par le Lieutenant Gouverneur, étant écrit, nous sommes assurés de la fidélité relative du discours rapporté dans cette publication.

2. À partir de 1964, le *Journal des débats* publie le verbatim des débats.

[1] La stratégie de recherche présentée ici correspond à une configuration particulière des outils disponibles dans le Portail ATO-MCD mettant en coopération trois composantes logicielles. On trouvera, dans le chapitre intitulé : « SATO_XML : une plateforme Internet ouverte pour l'analyse de texte assistée par ordinateur », la présentation générale de l'architecture informatique aux divers plans des données, des traitements informatiques et de l'infrastructure matérielle. Ce texte se concentre sur la ressource logicielle principale du projet ATO-MCD, soit le logiciel SATO-XML.

2. Principes pour un environnement coopératif

Le premier objectif d'un environnement coopératif de recherche est de permettre à plusieurs chercheurs de travailler simultanément au même projet de recherche. Pour qu'un tel environnement puisse remplir efficacement cet objectif, il doit offrir des instruments facilitant la coordination des actions des membres d'une même équipe. Afin de rencontrer cet objectif, l'ensemble des outils de traitement et d'analyse offerts s'articulent autour d'un Cahier de projet (Figure 1 en annexe). Les deux éléments principaux de celui-ci sont le Relevé des opérations et le Journal de bord. Le premier est le journal informatique des opérations effectuées sur le site par chacun des chercheurs du projet. Chacune des opérations produites par les chercheurs est consignée dans le Relevé des opérations. Quant au Journal de bord, il permet aux chercheurs de documenter leurs décisions stratégiques et d'en consigner la justification. Chacune des entrées du Journal de bord peut être reliée aux opérations informatiques afférentes du Relevé des opérations. Par exemple, le Journal de bord sera l'endroit où l'on pourra lire que le 10 novembre Guillaume a travaillé à la construction de la grille thématique. Cette entrée est liée aux opérations qui ont permis la construction de cette grille. On voit comment, en plus de favoriser la coordination, l'intégration autour d'un Cahier de projet favorise aussi la documentation et la *traçabilité* des opérations, indispensables pour la réussite d'une démarche de recherche intégrée. Le Cahier de projet est aussi le lieu où sont consignés l'ensemble des artefacts produits par les différents outils de traitement et d'analyse (AFC, tableaux statistiques, corpus, fragments annotés, etc.). Le cahier permet de garder une trace de tout le processus de recherche et de le reconstituer en tout temps. De plus, le moment venu, il permet de présenter l'ensemble du processus et de mieux cerner chacune de ses étapes.

C'est aussi pour permettre le travail coopératif de manière distribuée à plusieurs endroits qu'une interface web a été retenue. Ce choix possède plusieurs avantages pour le volet « diffusion » du projet. Enfin, l'interface web offre la possibilité d'utiliser le logiciel sur n'importe quelle plate-forme informatique (Mac, PC ou Linux).

Le second objectif de l'environnement coopératif est la « collaboration » informatique qui permet de fédérer, dans un environnement de recherche cohérent, un ensemble de composantes informatiques spécialisées. Dans cette première version, trois composantes spécialisées sont intégrées :

1) la composante SATO (http://nouvelle.ato.uqam.ca/) est utilisée pour tout ce qui relève du traitement textuel : établissement de sous-corpus contextuels, constitution de « tableaux lexicaux », catégorisation en contexte, etc. ;

2) la composante Guidexpert (http://fable.ato.uqam.ca/guidexpert/guidexpert.htm) est utilisée pour l'attribution de catégories grammaticales, la lemmatisation et certaines tâches de catégorisation thématique semi-automatiques ;

3) la composante R (http://www.r-project.org/) est utilisée pour tous les traitements statistiques (par exemple, les AFC).

L'intégration de ces composants est réalisée dans un système conçu à l'aide de l'environnement web Plone (http://www.plone.org/). Bien que chacune de ces trois composantes possède une interface propre qui demeure accessible à partir du Portail, l'idée centrale de la stratégie intégrée de recherche est de proposer une interface transversale qui intègre les spécialités de chacune des composantes à des instruments de documentation et de coordination du travail de recherche.

3. Construction et traitement des corpus

Sur le Portail, un corpus est l'ensemble des documents, enrichis des descriptions produites par l'application de grilles de catégories propres à chaque projet. Les chercheurs sont libres de définir leurs propres corpus et grilles analytiques ou de retenir des corpus et des grilles déjà existantes. Par exemple, le corpus CCC est composé d'un ensemble de documents reliés aux conférences constitutionnelles canadiennes et d'une grille de catégorie socio-sémantique. Ces documents et/ou cette grille pourraient être réutilisées dans un nouveau contexte de recherche. Dans la mesure où un chercheur désire constituer un nouveau corpus, il lui suffit d'envoyer vers le Portail la copie numérique des documents qu'il souhaite analyser.

Dans le cas du projet Ouverture, nous avons réutilisé des textes déjà disponibles dans le corpus (DNL) et nous avons numérisé les documents additionnels pour répondre à de nouvelles questions de recherche. Nous avons réuni ces documents en format *texte simple* (.txt) et nous les avons téléchargés sur le Portail dans le dossier que nous avions créé à cet effet. Lors du téléchargement des documents, en plus des informations de nature référentielle telles que le titre du document ou la référence bibliographique, certaines variables qui serviront à partitionner le corpus (par exemple, le moment de l'énonciation ou le type de locuteur) ont été spécifiées à l'aide de balises dans le texte numérisé. Dans le projet Ouverture, chaque document utilisait comme variable l'année, la session, la législature, le parti et le Premier ministre. Il aurait été possible d'ajouter d'autres variables et de cumuler plusieurs modalités d'une variable sur un même segment textuel. L'ensemble des variables ne recevront pas nécessairement de traitement statistique au moment de l'analyse. Enfin, des variables supplémentaires peuvent toujours être définies à une étape ultérieure de la recherche. Les outils servant à la construction de ces variables seront abordés dans la suite de l'exposé. L'ensemble des variables et de leurs modalités définies sur un corpus est accessible dans le Cahier de projet.

Les transformations dans la représentation informatique des corpus ainsi que les procédures d'indexation nécessaires pour leur traitement par les composantes Guidexpert ou SATO sont transparentes pour les chercheurs. Dans le cas où les fonctionnalités des différentes composantes utilisées par le Portail ne seraient pas jugées suffisantes, les chercheurs pourront exporter leur corpus dans un format qui permettra son traitement selon l'ensemble des fonctionnalités propres à chaque logiciel. En fait, à partir des informations fournies lors du téléchargement des documents du corpus, des versions du corpus adaptées aux différents logiciels sont automatiquement générées. Des versions des corpus utilisables directement dans SATO, Guidexpert ou d'autres logiciels fédérés sur le Portail sont toujours disponibles. Les usagers peuvent donc utiliser ces corpus adaptés aux logiciels de leur choix à leur gré pour réaliser des opérations supplémentaires.

Une fois le corpus complété, les chercheurs peuvent amorcer l'analyse en utilisant les divers outils pour construire différentes grilles descriptives ou produire des analyses statistiques à partir des tableaux générés à l'aide du système.

4. Outils pour la construction de grilles analytiques

À l'intérieur du Portail, une « grille descriptive» réfère à une représentation particulière des données lexicales ou textuelles sur laquelle des analyses statistiques pourront être effectuées. L'analyse la plus simple porte sur la distribution des formes lexicales dans différentes partitions du corpus. Il est possible cependant de travailler sur les formes lemmatisées produites à l'aide de Guidexpert. Les recherches du GRADIP ont plutôt privilégié les catégories socio-sémantiques appliquées aux noms communs et aux adjectifs (Duchastel et Armony, 1995).

Dans tous les cas, c'est la distribution des mots catégorisés ou non qui fera l'objet d'analyses statistiques.

La composante Guidexpert comporte des connaissances linguistiques et sémantiques du français et de l'anglais permettant de traiter une version lemmatisée du lexique et d'assigner automatiquement les fonctions syntaxiques aux formes lexicales. Cette description morpho-syntaxique peut-être utile pour limiter l'étendue de la catégorisation en contexte ou pour établir des sous-ensembles textuels, par exemple, pour une analyse de la distribution des adjectifs par locuteur. Dans le projet Ouverture, nous avons eu recours à Guidexpert afin de restreindre l'analyse aux noms et aux adjectifs. Nous avons utilisé les capacités linguistiques du logiciel Guidexpert pour assigner automatiquement une catégorie syntaxique à chaque mot du corpus. Par la suite, nous avons transféré ces informations dans la version SATO du corpus (Figure 2 en annexe).

Il n'existe pas de protocole universel de création ou d'application de ces grilles. Plusieurs approches existent. Une des possibilités est d'appliquer un dictionnaire de catégories (BDL, Banque de données lexicales) et de désambiguïser en contexte les lexèmes recevant plus d'une catégorie à l'aide de SATO. On peut aussi utiliser l'interface de Guidexpert pour explorer le corpus et construire de manière inductive une grille thématique à partir des champs sémantiques les plus fréquents. Pour ce faire, on crée une grille de thématisation qui associe à des modalités d'une variable thématique un ensemble de formes lexicales et de champs sémantiques larges ou restreints qui sont repérés automatiquement par Guidexpert. Quelle que soit la méthode utilisée pour construire cette thématisation, elle sera documentée dans le Cahier de projet et appliquée à partir du Portail. Cette thématisation pourra faire l'objet de traitements statistiques ultérieurs, si la variable répond aux propriétés requises. Dans la suite de l'exposé, nous ne reviendrons pas sur les fonctionnalités de thématisation sémantique de Guidexpert faute d'espace.

5. Outils pour l'analyse statistique

SATO permet de construire des tableaux représentant la distribution des unités lexicales (formes lexicales ou lemmatisées) ou d'une variable (par exemple, des catégories socio-sémantiques) à travers une partition donnée du corpus. Par exemple, on peut obtenir le tableau de la distribution des formes lexicales par locuteur ou le tableau de la distribution des catégories socio-sémantiques par période. Dans le cas du projet Ouverture, nous avons construit à l'aide de SATO le tableau de la distribution des lexèmes distingués selon leur fonction grammaticale par législatures (Figure 3 en annexe).

Le principe de partition que nous avons retenu pour ce tableau est la variable législature, c'est-à-dire la période d'exercice d'un gouvernement d'une élection à l'autre. La stricte chronologie annuelle ne permettait pas une analyse en fonction des partis ou des locuteurs, puisque, certaines années, deux Premiers ministres provenant de partis différents ont prononcé un discours d'ouverture. La variable Premier ministre ne permettait pas de distinguer les différents mandats d'un même Premier ministre. La variable Parti taisait l'aspect diachronique que nous privilégiions en analysant un corpus recouvrant plus de 40 ans. Évidemment, la variable législature nous permettait de retenir à la fois toutes les informations concernant le Premier ministre, le Parti et l'année.

Dans l'environnement SATO, ces tableaux peuvent être construits aussi bien sur l'ensemble du corpus que sur la distribution d'une unité ou d'une variable dans un sous-ensemble du corpus. La forme que peut prendre chaque tableau est variable. Par exemple, il est fréquent de s'intéresser à la distribution des « cooccurrences » ou, plus précisément, aux diverses formes

lexicales qui apparaissent dans le contexte (une ou plusieurs phrases, une borne numérique, etc.) d'une forme particulière.

Les tableaux peuvent être filtrés selon divers critères paramétrables par les chercheurs afin, par exemple, d'éliminer les hapax ou de ne retenir que les formes dont la fréquence totale est supérieure à la médiane. Tous ces tableaux peuvent être sauvegardés dans le Cahier de projet et peuvent être annotés. Dans le projet Ouverture, nous avons filtré le tableau en ne retenant que les noms et adjectifs dont la fréquence était supérieure à 10. Ainsi, nous avons expurgé les textes des mots outils pour ne garder que les mots présentant un fort contenu référentiel. En ce sens, nous sommes restés fidèles à la méthodologie du GRADIP : « Dans le cas particulier de la catégorisation socio-sémantique, telle que nous la concevons, on vise à classer – de manière exhaustive et exclusive – les mots à valence référentielle (noms et adjectifs) en fonction d'un système de catégories thématiques. » (Duchastel et Armony, 1995). En ne retenant que les noms et les adjectifs présentant plus de dix occurrences dans l'ensemble du corpus, nous sommes passés de 13034 lexèmes à 1042 lexèmes. Au terme de toutes ces manipulations, nous traitons 74 % des adjectifs et des noms communs, soit 19% du corpus total.

L'instrument d'analyse statistique privilégié pour ces volumineux tableaux est l'analyse factorielle des correspondances (Lebart et Salem, 1994). Cette analyse permet de visualiser les variations dans la distribution des unités entre les parties du corpus (locuteurs, périodes, etc.) Outre les vertus heuristiques de cette visualisation, le choix privilégié de cette méthode d'analyse se justifie par ses multiples propriétés qui la rendent robuste vis-à-vis des perturbations dans les données. (Viprey 2003 ; *Lebart et al.*, 2000). Pour le projet Ouverture, cette méthode était toute indiquée puisque les discours varient beaucoup en taille. Les AFC ont été réalisées à l'aide de R sur le tableau construit précédemment.

Le Portail offre plusieurs outils pour l'aide à l'interprétation des AFC. Outre la possibilité de visualiser dans un même plan deux axes de l'AFC (on s'intéressera la plupart du temps aux diverses combinaisons des premiers axes), plusieurs fonctionnalités permettent de pallier la difficulté de visualiser la projection de plusieurs centaines d'éléments dans le plan. Par exemple, on pourra limiter la représentation des points-lignes à celles qui répondent à certains critères comme la fréquence. Pour affiner l'interprétation, on pourra aussi afficher dans une teinte différente les éléments qui sont particulièrement bien représentés dans le plan (seuil paramétrable applicable au cosinus carré) ou ceux qui ont une importante contribution dans la construction d'un des axes du plan. Ce sont ces stratégies que nous avons adoptées pour le projet Ouverture. Nous avons limité l'affichage des points-lignes aux « meilleurs » de chacun des axes projetés dans le plan, c'est-à-dire les points lignes ayant la meilleur position sur l'axe (COS2) jusqu'à ce que l'ensemble sélectionné représente 60% des contributions à l'axe. Dans le graphique présenté en annexe (Figure 4), les points-lignes affichés sont l'union des ensembles liés à chaque axe projeté sur le plan.

Encore une fois, les différentes visualisations produites peuvent être sauvegardées dans le Cahier de projet. Comme tous les éléments sauvegardés dans le Cahier de projet, la visualisation particulière reste aussi liée aux objets dont elle est issue afin de pouvoir retracer l'origine de sa construction**.**

6. Conclusion

Le Portail ATO-MCD se veut un environnement permettant la recherche coopérative. L'élaboration de la stratégie de recherche proposée est axée sur l'intégration des différents outils de construction des corpus, de traitement des données et d'analyse statistique autour d'un Cahier de projet qui permet la coordination, mais surtout la documentation du processus

de recherche. En ce sens, le Portail propose de rendre possible et/ou de faciliter l'application de méthodologies « courantes » dans le domaine de l'analyse de texte par ordinateur. Les concepts, méthodes et outils qu'il rend ainsi accessibles sont bien connus des praticiens de l'ATO. Ainsi, le Portail est un environnement qui intègre ces méthodes éprouvées de manière cohérente, favorisant ainsi le développement d'un « communauté argumentative » de chercheurs en ATO. De cette volonté découle le choix d'organiser l'interface usager autour du Cahier de projet, qui facilite la documentation et la justification des décisions inhérentes à tout processus de recherche. Le Portail permet donc le développement d'outils de coopération sur deux plans. Tout d'abord, il permet la coopération et le suivi des opérations par plusieurs chercheurs sur un même projet. Ensuite, il permet l'intégration des fonctionnalités de plusieurs logiciels d'analyse de texte et de traitement des données.

Dans l'avenir, les développements du Portail porteront sur l'intégration d'autres types d'outils analytiques éprouvés. La priorité sera donnée à l'intégration d'analyses statistiques qui complémenteront l'AFC. La classification hiérarchique et le calcul des spécificités sont des exemples d'instruments statistiques qui pourraient enrichir l'éventail des outils disponibles sur le Portail (Lebart et Salem, 1994). Une attention particulière sera aussi portée à l'intégration des méthodes pouvant être utilisées sur des structures de graphes pour l'étude des cooccurrences (Batagelj et Mrvar, 2003).

Références

Batagelj V. et Mrvar A. (2003). Developing Pajek - Exploratory analysis of networks. Analyse des Données Relationnelles / EHESS-INED. INED. http://vlado.fmf.uni-lj.si/pub/networks/doc/seminar/paris03.pdf.

Duchastel J. et Armony V. (1995). La catégorisation socio-sémantique. In *Actes des JADT 1995 :*193-200.

Lebart L. et Salem A.. (1994). *Statistique textuelle*. Dunod.

Lebart L., Morineau A. et Piron M. (2000) *Statistique exploratoire multidimensionnelle. 3e édition.* Dunod.

Viprey J.-M. (2003). *Morneille, Colière et messieurs Labbé*. http://laseldi.univ-fcomte.fr/morneille.htm

Viprey J.-M. (2002) : *Analyses textuelles et hypertextuelles des Fleurs du Mal* [texte intégral et moteur de recherche sur CD-Rom; exploration lexicale, morpho-syntaxique, prosodique, phonématique], Champion, Lettres Numériques n°5.

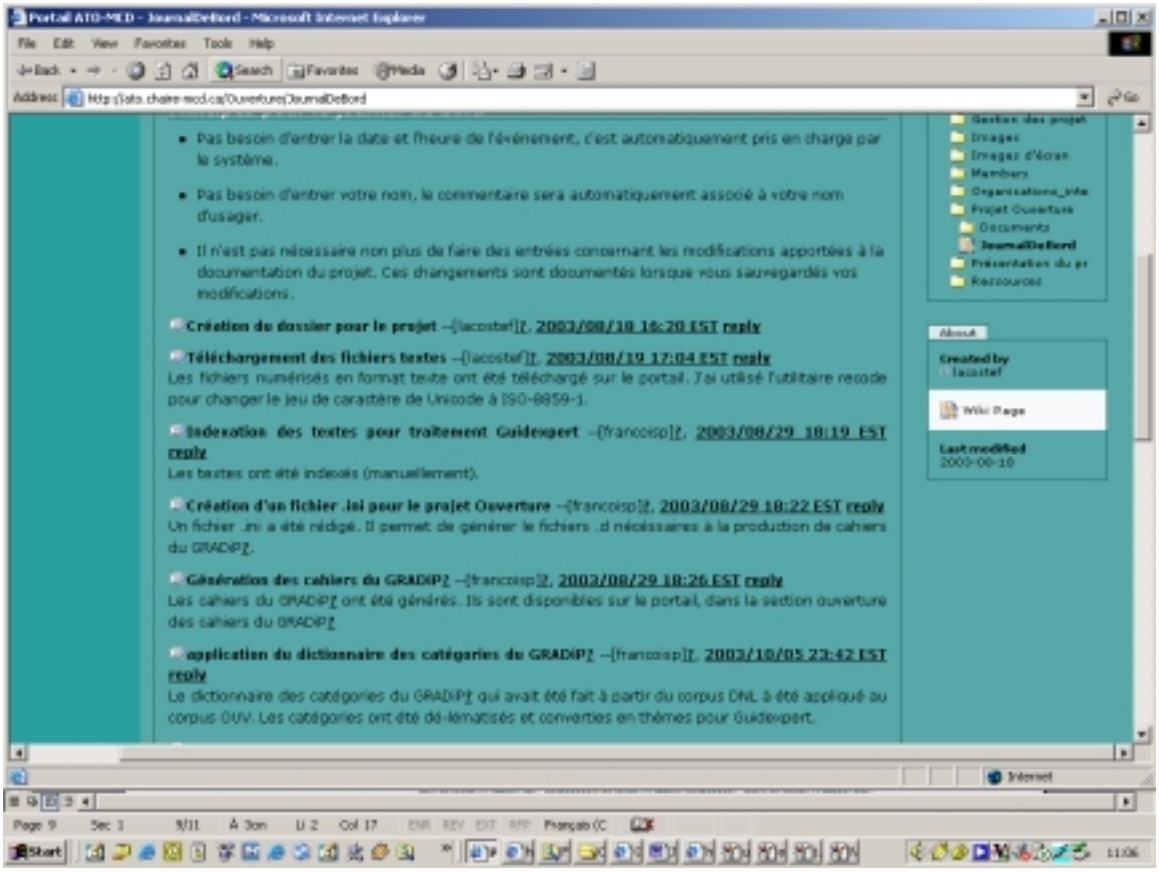


Figure 1. Cahier de Projet

```
1  *en=al960¶
2  *legis=26¶
3  *parti=plq¶
4  *pm=lesage¶
5  *session=2¶
6  *Phrase=3 Première^Cs=|adj|^Lemme="premier" tâche^Cs=|nc|^Lemme="tâche" du^Cs=|prep|^Lemme="du"
   gouvernement^Cs=|nc|^Lemme="gouvernement" a été de^Cs=|prep|^Lemme="de" déterminer^Cs=|vinf|^Lemme="déterminer" les
   besoins^Cs=|nc|^Lemme="besoin" les plus urgents^Cs=|adj|^Lemme="urgent"¶
7  de^Cs=|prep|^Lemme="de" la province^Cs=|nc|^Lemme="province".^Cs=|ponct|^Lemme="." *Phrase=4 Il a
   déjà^Cs=|adv|^Lemme="déjà" adopté^Cs=|vppa|^Lemme="adopter" des^Cs=|prep|^Lemme="des"
   initiatives^Cs=|nc|^Lemme="initiative" concrètes^Cs=|adj|^Lemme="concret".^Cs=|ponct|^Lemme="." *Phrase=5
   Au^Cs=|prep|^Lemme="au" cours^Cs=|prep|^Lemme="cours" de^Cs=|prep|^Lemme="de" la¶
8  session^Cs=|nc|^Lemme="session" présente^Cs=|adj|^Lemme="présent", il se propose^Cs=|vfle|^Lemme="proposer"
   de^Cs=|prep|^Lemme="de" soumettre^Cs=|vinf|^Lemme="soumettre" aux^Cs=|prep|^Lemme="aux"
   chambres^Cs=|nc|^Lemme="chambre" un^Cs=|num|^Lemme="un" programme^Cs=|nc|^Lemme="programme" de^Cs=|prep|^Lemme="de"¶
9  législation^Cs=|nc|^Lemme="législation" visant^Cs=|vppa|^Lemme="viser" à^Cs=|prep|^Lemme="à"
   répondre^Cs=|vinf|^Lemme="répondre" aux^Cs=|prep|^Lemme="aux" exigences^Cs=|nc|^Lemme="exigence"
   collectives^Cs=|adj|^Lemme="collectif" les plus pressantes^Cs=|adj|^Lemme="pressant" de^Cs=|prep|^Lemme="de" la¶
10 population^Cs=|nc|^Lemme="population", à^Cs=|prep|^Lemme="à" élargir^Cs=|vinf|^Lemme="élargir" le
   champ^Cs=|nc|^Lemme="champ" d'^Cs=|prep|^Lemme="d'"action^Cs=|nc|^Lemme="action" et
   accroître^Cs=|vinf|^Lemme="accroître" l'efficacité^Cs=|nc|^Lemme="efficacité" du^Cs=|prep|^Lemme="du"
   gouvernement^Cs=|nc|^Lemme="gouvernement"¶
11 par^Cs=|prep|^Lemme="par" la création^Cs=|nc|^Lemme="création" de^Cs=|prep|^Lemme="de"
   nouveaux^Cs=|adj|^Lemme="nouveau" ministères^Cs=|nc|^Lemme="ministère", à^Cs=|prep|^Lemme="à"
   moderniser^Cs=|vinf|^Lemme="moderniser" ou remodeler^Cs=|vinf|^Lemme="remodeler" l'appareil^Cs=|nc|^Lemme="appareil"¶
12 administratif^Cs=|adj|^Lemme="administratif" existant^Cs=|vppa|^Lemme="exister".^Cs=|ponct|^Lemme="." *Phrase=6 Le
   gouvernement^Cs=|nc|^Lemme="gouvernement" vous invitera^Cs=|vfle|^Lemme="inviter" à^Cs=|prep|^Lemme="à"
   étudier^Cs=|vinf|^Lemme="étudier" un^Cs=|num|^Lemme="un" projet^Cs=|nc|^Lemme="projet" de^Cs=|prep|^Lemme="de"
   loi^Cs=|nc|^Lemme="loi"¶
13 pour^Cs=|prep|^Lemme="pour" autoriser^Cs=|vinf|^Lemme="autoriser" la création^Cs=|nc|^Lemme="création"
   d'^Cs=|prep|^Lemme="d'"un^Cs=|num|^Lemme="un" ministère^Cs=|nc|^Lemme="ministère" des^Cs=|prep|^Lemme="des"
   \Affaires^Cs=|np|^Lemme="Affaires" culturelles^Cs=|adj|^Lemme="culturel" qui aura pour^Cs=|prep|^Lemme="pour"¶
14 sa juridiction^Cs=|nc|^Lemme="juridiction", entre autres organismes^Cs=|nc|^Lemme="organisme", un^Cs=|num|^Lemme="un"
   \Office^Cs=|np|^Lemme="Office" de^Cs=|prep|^Lemme="de" la linguistique^Cs=|nc|^Lemme="linguistique",
   un^Cs=|num|^Lemme="un"¶
15 département^Cs=|nc|^Lemme="département" du^Cs=|prep|^Lemme="du" \Canada^Cs=|np|^Lemme="Canada"
```

Figure 2. Transfert des informations sémantiques et lemmatiques de Guidexpert vers SATO

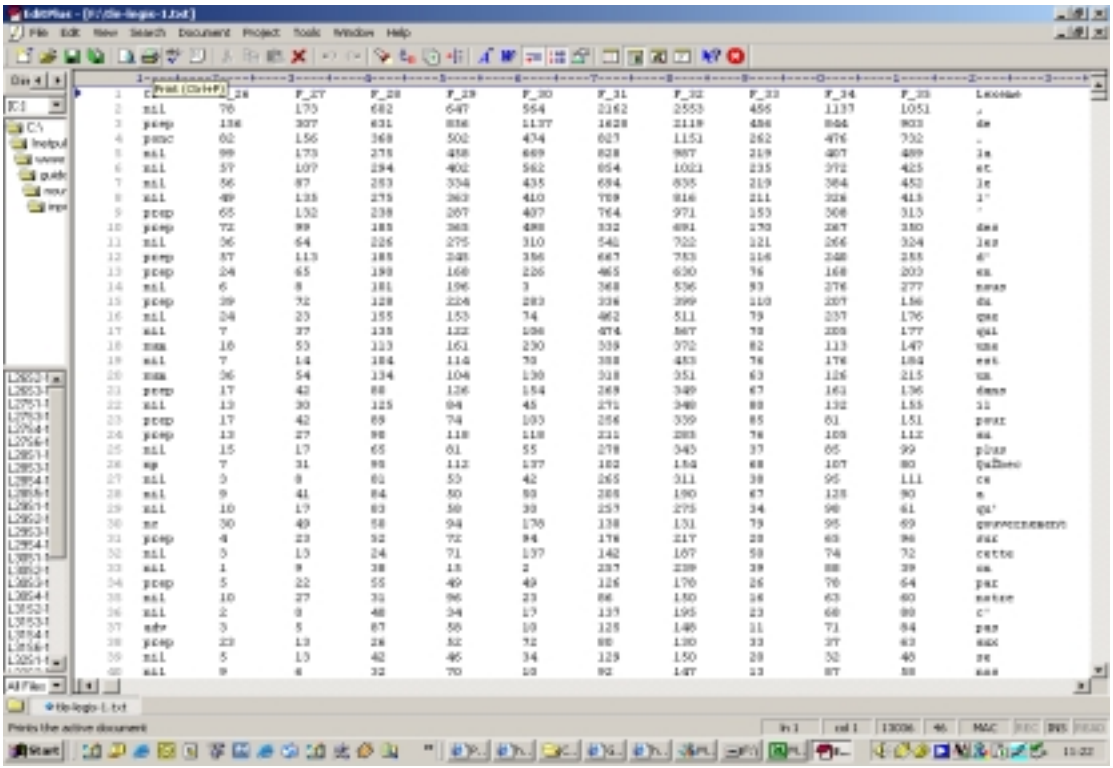


Figure 3. Tableau lexical entier des noms et adjectifs présentant plus de dix occurrences

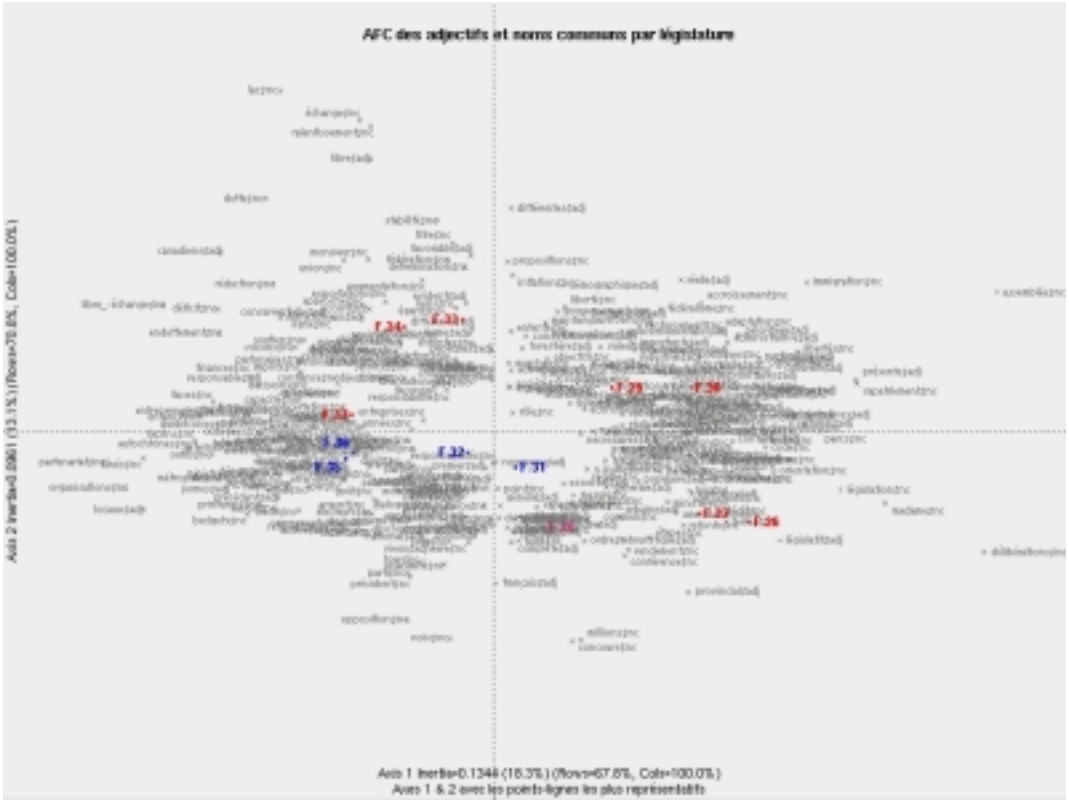


Figure 4. Analyse factorielle des correspondances produite à partir du tableau lexical entier

Le discours de la BCE concernant les aspects sociaux

Anne Dufresne

GRAID – Institut de sociologie – Université Libre de Bruxelles
Campus du Solbosch – Bruxelles – Belgique
dufresne@ose.be

Abstract

This article emphasizes the discourse of the European Central Bank (ECB), an extremely revealing insight into the EU ideology, which has acquired a political weight and influence over the running of economic and social affairs. The lexical study and the analysis of the main lexical and discursive strategies found in the monthly bulletin of the ECB over the 1999-2003 period will show how, beyond its major aim of monetary stability, the Bank is also very interested in wage and labour market issues. More specifically, we'll examine the transformation of the vocabulary built around the wage issue. For instance, we will see how the Bank uses arguments about macro-economic constraints as pretexts to obtain more "responsible", i.e. "moderate" trade unions' claims.

Résumé

Cet article met en évidence le discours de la Banque centrale européenne (BCE), révélateur extrême de l'idéologie de l'Union européenne (UE), qui a acquis un poids politique et une influence sur la conduite des affaires économiques et sociales. L'étude du lexique et l'analyse des stratégies lexicales et discursives sur les bulletins mensuels de la Banque entre 1999 et 2003 montre comment, au delà de son objectif principal de stabilité monétaire, elle s'intéresse aussi, et de très près, aux questions salariales et de marché du travail. Cette étude lexicale s'attache plus particulièrement à décrire la transformation du vocabulaire autour de la question salariale. Par exemple, nous verrons comment la banque utilise des arguments de contrainte macro-économiques comme prétexte pour « responsabiliser », i.e. « modérer », les revendications des syndicat

Mots-clés : idéologie communautaire (i.e. de l'Union européenne), doxa néo-libérale, Banque centrale européenne, modération salariale, stabilité monétaire, prétextes macro-économiques, partenaires sociaux, marchés financiers, analyse statistique textuelle de corpus.

Keywords: EU ideology, neoliberal doxa, European Central Bank, wages moderation, monetary stability, macroeconomic pretexts, social partners, financial markets, statistical analysis of the textual corpus.

1. Introduction

Dans une récente communication sur l'analyse des Grandes orientations de politiques économiques (GOPE)[1], Corinne Gobin[2] (2003) évoque « *le "changement de régime politique" imposé par l'union économique et monétaire (UEM) qui se met en oeuvre à travers la transformation du sens des mots et la diffusion d'un nouveau système de représentations collectives* ». Ce nouveau système répond, dans le champ des économistes, au « tournant théorique » qui s'est produit dans les années 1970 et 1980. En effet, le monétarisme puis la nouvelle macroéconomie classique ont contribué à l'émergence d'un corpus doctrinal intégré dont les préceptes sont devenus incontournables (Raveaud, 2003). De multiples « nécessités » sont

[1] Les GOPE contiennent les consignes à suivre pour que les Etats membres respectent les engagements économiques et monétaires contenus dans le Traité de Maastricht (critères de convergence), puis dans le Pacte de stabilité et de croissance.

[2] Je remercie chaleureusement Corinne Gobin pour son aide précieuse tant pour la lexicométrie que pour ses commentaires.

alors apparues : la « flexibilité » du « marché du travail », de l'augmentation du « taux d'emploi » (et non la réduction du taux de chômage), ou bien encore la « modération salariale » par la « responsabilisation » des « partenaires » sociaux. Ces notions seront détaillées plus loin, l'idée étant de déconstruire ce jargon communautaire pour pouvoir construire de nouvelles catégories et concepts.

Pour révéler ce corpus doctrinal intégré évoqué, il nous a semblé particulièrement pertinent de décrypter les discours de la Banque centrale européenne (BCE) qui a acquis un poids politique et une influence sur la conduite des affaires économiques et sociales et dont le discours est un révélateur extrême de l'idéologie de l'UE. Nous tenterons ainsi de montrer, par l'étude du lexique et l'analyse des stratégies lexicales et discursives sur les bulletins mensuels de la Banque comment, au delà de son objectif principal de stabilité monétaire, elle s'intéresse aussi, et de très près, aux questions salariales et de marché du travail.

Préliminaires : Méthodologie et Présentation du corpus

Voyons tout d'abord comment la BCE agence son discours dans ses bulletins pour mieux comprendre son argumentaire. Le corpus de textes sur lequel nous opérons les recherches est constitué par la réunion des bulletins mensuels de la BCE sur 4 ans : du 1er janvier 1999 au 31 décembre 2002. La date du 1er janvier 1999 correspond à l'année où entre en vigueur le traité d'Amsterdam et où se met en route le système européen de banques centrales (SEBC) et le lancement de la monnaie unique. Constitué de 48 textes, il comprend 1373695 occurrences et 13506 formes lexicales[3]. On peut constater (cf. tableau 1 ci-dessous) une légère augmentation en volume des bulletins entre 1999 et 2002.

Partie	*Nb occurrences*	*Nb formes*
1999	321625	8001
2000	315728	7745
2001	367985	8665
2002	368357	8574

Tableau 1 : Répartition par année des formes lexicales dans les parties du corpus

Le bulletin représente l'un des moyens de communication écrit de la BCE les plus importants. Sa publication répond à l'obligation pour la BCE de publier au moins chaque trimestre des rapports sur les activités du SEBC. Elle y explique au public les décisions de politique monétaire prises par le Conseil des gouverneurs.

La structure des Bulletins se retrouve quasiment identique de mois en mois, l'« *éditorial* » résume les motivations qui sous-tendent les décisions de politique monétaire. Ces questions sont ensuite exposées plus en détail dans la partie intitulée « *Les évolutions économiques de la zone euro* ». La troisième composante-type du Bulletin intitulée « *Statistiques de la zone euro* » comprend une série de tableaux et graphiques (qui ont été supprimé pour la constitution du corpus, tout comme les tableaux et graphiques apparaissant dans les autres parties). Sont également publiés dans chaque bulletin un ou deux articles sur des questions diverses

[3] Le logiciel d'analyse statistique lexicale que nous utilisons, Lexico3, a été mis au point par le laboratoire du SYLED (Université de Paris 3) sous la direction d'André Salem. Suivant le mode de segmentation des textes utilisé par ce logiciel, une forme lexicale correspond à une suite particulière de caractères graphiques séparés par des caractères délimiteurs (espaces et signes de ponctuation), ce qui fait que chaque variation grammaticale (verbes, adjectifs, substantifs) correspond à une forme lexicale différente (marché, marchés, …), les occurrences étant le nombre d'attestations rencontrées de ces formes.

liées à la conception et à la conduite de la politique monétaire ainsi que sur les autres points importants pour la BCE. En janvier 1999, le premier Bulletin explique que les suivants traiteront « particulièrement des facteurs dont dépend le succès de la politique monétaire et la stabilité de l'euro : notamment de sujets comme les évolutions des salaires et des marchés du travail, les rigidités structurelles, et les politiques budgétaires[4] ». Dans cet article, nous nous intéressons en particulier à la place et au poids relatif dans l'ensemble du corpus de la thématique du « marché du travail » qui inclut la question du « coût du travail ». Nous n'analysons pas ici l'évolution des interventions ou de la stratégie de la BCE sur les marchés monétaire ou de change[5].

2. « La » mission monétaire sans objectifs économiques

Pour entrer dans l'idéologie libérale bancaire, ce chapitre analyse le positionnement adopté par la BCE dans ses bulletins mensuels sur les questions macro-économiques, au regard de « sa » mission principale de stabilité monétaire. Dépassant les postulats habituels de stabilité des taux et d'amélioration des soldes budgétaires, nous nous focalisons sur les dangereuses priorités politiques qui en émergent.

Nous rappelons ici les principes fondant le pouvoir monétaire pour mieux comprendre la teneur du vocabulaire bancaire. La BCE est le gardien vigilant de l'orthodoxie monétaire. Elle est indépendante et met en œuvre la politique monétaire unique dans les douze États de la zone euro. Conformément à l'article 105 du traité, son objectif principal est de maintenir la stabilité des prix[6]. Toutefois, sans préjudice de cet objectif, elle est aussi tenue par le traité de mettre en œuvre une politique monétaire soutenant « les politiques économiques générales », en vue de contribuer à la réalisation des objectifs économiques.

Par l'analyse des bulletins, on a pu constaté que les objectifs économiques de croissance et d'emploi ont été oubliés par la BCE, au profit de l'objectif central de la stabilité monétaire. C'est pourquoi, on s'étonne dans un premier temps de la place réservée à la forme « ***croissance*** » (5e rang dans l'index hiérarchique des fréquences[7], 5340 occ.) dans le corpus. Cette importance est en fait très relative dans la mesure où « croissance » peut certes être entendue comme variable économique en tant que telle, mais aussi comme substantif synonyme d'augmentation. On s'aperçoit alors que la famille lexicale liée à la croissance de la production[8] ne représente environ que 16 % de l'usage de croissance, tandis que domine la famille liée à la « *croissance monétaire et financière* ». Quant à la « *croissance de l'emploi* », elle apparaît pour 7 % de l'usage de croissance. La forme lexicale « emploi », en outre, s'associe relativement fréquemment avec celle de « croissance ». Contrairement à ce que révèle l'analyse des GOPE, on constate que c'est le cas surtout dans l'extension droite de cette forme. Etant donnée la logique monétariste de la BCE, la croissance est une variable sur laquelle la BCE semble avoir renoncé à agir. Dans l'univers macro-économique contraint décrit précé-

[4] D'autres études traitent en détail de la position de la BCE. Concernant le marché du travail, cf. Dufresne (2000a et 2000c) ; sur la politique budgétaire, cf. Dufresne (2000b) ; sur les retraites, cf. Dufresne (2001c). Ces textes sont disponibles sur le site de l'Observatoire social européen (www.ose.be).

[5] Pour plus de détails sur ces questions Cf. Artis (2003) ou Artus et Wyplosz (2002).

[6] La BCE a quantifié la stabilité comme étant une hausse de l'indice des prix à la consommation harmonisé de moins de 2 % sur le moyen terme.

[7] Cette liste de formes rangées par ordre de fréquence a été obtenue en ne conservant de l'index hiérarchique tel que issu du logiciel Lexico-3 que les noms, les adjectifs, les adverbes et les verbes.

[8] La famille lexicale liée à la production comprend la *croissance économique* (297 occ.), la *croissance du PIB* (403 occ.), et la *croissance de la production* (203 occ.).

demment, les actions se réduisent à enrichir le contenu en emplois de la croissance en particulier par une réduction du coût du travail. Dans les GOPE, la forme lexicale « emploi » s'emploie prioritairement à la gauche de croissance mais est attestée à gauche aussi de façon non négligeable. « *La croissance doit se traduire en des créations d'emplois mais le taux d'emploi doit permettre la croissance. Ce cercle vertueux à mettre en place est ainsi d'abord mis en scène à travers ce* ***ballet lexical circulaire*** *qui fait qu'au bout du compte, par l'effet de la répétition, on ne sait plus très bien quel est l'objectif : la croissance pour l'emploi ou l'emploi pour la croissance*. » (Gobin, 2003).

Rappelons maintenant brièvement la logique globale de la Banque pour mieux comprendre la « nécessité » des réformes structurelles sur le « marché du travail » : son but est d'atteindre une réelle discipline[9] (budgétaire et salariale) pour réduire le risque de tensions inflationnistes à court terme, et pouvoir mener à moyen terme une politique monétaire moins restrictive qui, en maintenant les taux d'intérêt faibles favorise l'investissement, la croissance et donc l'emploi. Selon la banque, seules des « réformes structurelles sur le marché du travail » ainsi qu'une concurrence accrue sur les autres marchés doivent accompagner cette politique monétaire restrictive. Les nombreuses citations explicitent cette logique en trois temps :

1. La politique économique est réputée neutre. La BCE estime que :

   « *Les politiques économiques pourraient aussi avoir une influence sur la possibilité de parvenir à un retournement de la tendance générale à long terme du chômage : la politique budgétaire en mettant l'accent sur le ralentissement des dépenses plutôt que sur l'alourdissement de la fiscalité ; la politique monétaire en maintenant son objectif de stabilité des prix de manière durable et crédible*. » (rapport annuel, 1999)

Elle insiste en particulier sur la politique monétaire qui ne saurait être que pro-cyclique afin de ne pas fausser les anticipations réelles des agents.

> « *Vouloir réduire le chômage en mettant en œuvre une politique monétaire inflationniste serait, au bout du compte, voué à l'échec ; en effet, une telle stratégie ne ferait que nuire à la stabilité des prix à moyen terme, qui constitue le fondement d'une croissance durable de* l'emploi. » (janvier 1999)

> « *La politique monétaire ne peut traiter les problèmes structurels, qui trouvent leur origine dans d'autres domaines de la politique économique. La persistance de niveaux élevés de chômage signifie qu'il convient de s'attaquer résolument aux problèmes structurels importants qui se posent sur les marchés des biens et du travail*. » (mai 1999)

2. Si elle refuse d'accorder un rôle actif à la politique économique dans la création d'emplois, la Banque favorise systématiquement l'idée des réformes structurelles sur le marché du travail pour réduire le chômage :

   « *Dans la situation actuelle, le défi majeur que doit affronter la zone euro demeure la recherche de solutions aux problèmes structurels qui sont la cause principale des niveaux élevés de chômage*. » (avril 1999)

   « *On affirme souvent que des chocs macroéconomiques, en interaction avec l'organisation du marché du travail, ont également pu contribuer à l'augmentation du chômage en Europe. Des cadres institutionnels exagérément rigides ont peut-être constitué un frein aux ajustements imposés par les changements de l'environnement économique, d'où la persistance d'un chômage plus important*. » (mai 2000)

[9] Sur les 51 occurrences de discipline dans l'ensemble des bulletins, il est à noter que 30 font référence à la discipline budgétaire.

3. Il est également important de noter, que, en revanche, elle souhaite prendre en compte les réformes structurelles dans sa stratégie monétaire :

« *Des politiques structurelles efficaces à l'échelle de la zone euro conduiraient à une croissance réelle tendancielle plus élevée que l'eurosystème, dans le cadre de sa stratégie de politique monétaire, prendrait évidemment en compte*. »

Des taux et des temps : Le monde d'incertitudes de la sphère financière

L'examen de la liste des formes lexicales les plus fréquentes nous a permis de mettre en évidence le vocabulaire de base de la Banque, d'une part, on y voit l'importance du lexique lié à « la **quantification** » : *% et taux* (1er et 2e rang de la liste), *euro* (3e rang), *prix* (6e rang), *cours* (11e rang), *monétaire* (12e rang), *intérêt* (13e rang), *opérations* (18e rang), *milliards (*28e rang) ; et d'autre part celui du vocabulaire lié au **temps** et à la périodicité : *terme* (8e rang), *mois* (9e rang) , *trimestre* (14e rang) *période* (30e rang). On ne s'étonnera pas dans un tel corpus que les principales recommandations bancaires dirigées par une stratégie monétariste, favorisent les variables monétaires du fait de l'obsession de la stabilité des taux, mais il est plus étonnant de constater la contamination de ce vocabulaire quantitatif aux analyses de la banque concernant le « marché du travail ». Elle surestime systématiquement les variables monétaires et néglige systématiquement les données institutionnelles.

Ce vocabulaire s'adresse en particulier aux marchés financiers, qui doivent valider par des chiffres et sur des périodes données la politique d'orthodoxie monétaire (et budgétaire) décrite ci-dessus, seule norme crédible de politique économique à leurs yeux. C'est donc bien cet « impératif » de crédibilité vis-à-vis des marchés financiers, qui pousse la BCE à consolider sa politique monétariste et à se référer en permanence au monde d'incertitudes qui caractérise les marchés financiers. En effet, la famille lexicale de l'incertitude est très développée dans tout le corpus : la « *confiance* » (701 occ.) apparaît comme une variable essentielle qu'il s'agit de consolider, de maintenir ou d'améliorer. En outre, « *peut-être* » (467 occ.), « *probablement* » (362 occ.), « *sans doute* » (166 occ.), lexèmes exprimant l'incertitude, reviennent très fréquemment dans l'analyse du « marché du travail ». Par exemple :

« *Les effets néfastes des « coûts élevés de la main-d'œuvre » - mais aussi d'un salaire minimum - se concentrent sans doute sur certaines catégories de personnel comme les jeunes ou les peu qualifiés dont la productivité est peut-être inférieure à la* moyenne. » (novembre 2002).

Nous avons également été étonnés de découvrir dans la liste des 40 formes les plus fréquentes du corpus que, en dehors du vocabulaire déjà évoqué des « taux et des temps » et à l'exception de la croissance ne ressortent alors plus que deux termes essentiels : *marché* (16e rang) et *marchés* (39e rang) ainsi que *évolution* (20e rang) et *évolutions* (27e rang). Ces termes conduisent à une vision désincarnée de l'orientation économique comme si le marché avait une dynamique propre en terme de mouvement.

Dans ce contexte, on peut aussi remarquer le poids du verbe "devrait" ou « devraient » (1400 occ.). Tout comme pour les GOPE, « *un examen de son usage montre que nous sommes globalement dans le registre de l'injonction déguisée (adoucie par le conditionnel) ou de la « prédiction » économique, où l'on annonce comme une tendance ce que l'on voudrait obtenir. Son usage s'inscrit dans une stratégie discursive qui permet d'escamoter la question de l'acteur politique : ou l'Etat et les acteurs sociaux qui sont désignés dans ce discours sont invités à suivre une orientation précise ou, et c'est le cas le plus fréquent, ce qui est sujet du verbe « devrait » est une orientation économique désincarnée* » (Gobin, 2003).

L'impasse faite sur le rôle de la sphère politique dans le choix des options économiques et l'omniprésence de la « loi des marchés » se retrouve de la même manière dans le discours sur les « évolutions de salaires ».

3. La modération salariale

La question salariale ne fait pas partie du mandat de la BCE au sens strict. Pourtant, la division de la banque intitulée « *Euro Area Macro* » qui comprend deux unités : « *Ouput Demand & Labour Markets* » et « *Prices and Costs* » travaillent de près sur l'évolution conjoncturelle des salaires en Europe[10] sous prétexte de leur impact économique sur la consommation et l'inflation. Selon la Banque, « *il est fondamental de suivre attentivement l' évolution des salaires et de disposer de données fiables avec une périodicité élevée[…] mensuelle ou trimestrielle* » pour deux raisons essentielles : la stabilité monétaire qui correspond à la classique « spirale inflationniste » et le maintien de la compétitivité.

Selon la logique déjà évoquée, c'est bien la politique monétaire qui prime sur les autres politiques économiques, et en particulier sur la politique salariale qui devient une variable d'ajustement. C'est pourquoi, on ne parle pas de politique de revenu ni de négociations collectives au niveau communautaire, mais bien d'« *évolution des salaires* ». La BCE indiquerait-elle aux partenaires sociaux l'évolution des salaires tolérées ? Potentiellement habilitée à demander des objectifs salariaux aux syndicats pour la zone Euro dans l'enceinte du dialogue macro-économique[11], elle devrait déterminer également ce qu'elle considère être une norme acceptable ou pas. C'est pourquoi, depuis septembre 2000, la BCE a entamé un travail de recherche en collaboration avec Eurostat sur des « indicateurs de coût du travail dans la zone euro ».

Chaque mois, dans le tableau intitulé « *Evolution des prix et des coûts dans la zone euro* » sont mentionnés dans la rubrique « *autres indicateurs de coûts et de prix* » : les coûts unitaires de main d'œuvre, le revenu par personne occupée, et la productivité du travail. Plus généralement, dans l'ensemble du corpus, on constate que la forme « salaires » apparaît essentiellement pour faire remarquer le risque de leur augmentation par de multiples expressions[12], qui correspondent à plus d'un tiers de l'usage de « *salaires* » (112 occ. sur les 309 occ.).

D'autres syntagmes récurrents s'y ajoutent, comme les « *tensions sur les salaires* », ainsi que la « *modération salariale* » (67 occ.). Les salaires semblent donc « *évolués* » d'eux-mêmes plus qu'être « *négociés* »[13] par des « *partenaires sociaux* » (20 occ.) qui n'apparaissent que très peu, et seulement pour être « responsabilisés » par la Banque. Le salaire apparaît exclusivement comme un coût. La famille des « *coûts salariaux* » (274 occ.) représente à elle seule plus de 30 % de l'usage de « *coûts* ». On distingue alors deux grandes sous-famille : les « *coûts salariaux unitaires* » (ou « *coûts unitaires de main d'œuvre* ») sont les plus employés

[10] Une conférence, organisée par « l'*ECB Labour Market Workshop* » intitulée « *How are Wages determined in Europe?* » a eu lieu le 10-11 décembre 2001 à la BCE. Un livre reprenant les actes du colloque sera bientôt publié.

[11] Le dialogue macro-économique confronte la BCE aux questions liées à la politique salariale. Créé lors du Sommet de Cologne en juin 1999, il instaure un dialogue entre autorités budgétaires, responsables de la formation des salaires et de la politique monétaire. Lors de la première prise de contact, le Président en exercice du Conseil a déclaré qu'il espérait que le dialogue fera naître « une prise de conscience de la responsabilité conjointe des participants dans la gestion des paramètres macro-économiques ».

[12] Croissance des salaires (11 occ.), hausse des salaires (12 occ.), augmentations de salaires (10 occ.), hausses de salaire (13 occ), croissance des salaires (13 occ.), hausse de salaires (14 oc.) ; hausses de salaires (23 occ.), augmentations de salaires (16 occ.)

[13] « Evolutions des salaires » (22 occ.), « évolutions salariales » (33 occ.) s'opposent à « négociations salariales »(27 occ.) et « salaires négociés » (11 occ.).

(avec 64 % de l'usage de « *coûts salariaux* »), le restant correspondant aux « *coûts salariaux horaires* » (ou « *coûts horaires de main d'œuvre* »). En revanche, les « *rémunérations par personne occupée* » (10 occ.) ou *« par tête* » (34 occ.) n'apparaissent que très peu. On constate ainsi que ce sont bien les « *coûts salariaux unitaires* » censés mesurer les salaires corrigés de la productivité (c'est-à-dire la bonne mesure du risque inflationniste) qui se substituent au terme « salaires ».

Après avoir étudié son vocabulaire en matière salariale, il est intéressant de noter comment la BCE développe un argumentaire paradoxal en trois temps dans l'éditorial de presque tous les bulletins. Tout d'abord, elle se félicite de la « bonne modération » du ou des mois passés (trimestre ou semestre) tout en prévoyant, quelques lignes plus tard un risque d'augmentation pour l'avenir, avant de terminer par un satisfecit global pour maintenir la confiance des investisseurs. Elle ménage la chèvre (les investisseurs) et le chou (les partenaires sociaux)

Encadré : Argumentaire paradoxal sur la modération salariale

Exemple du bulletin de février 1999 :

Etape 1 : Félicitations aux partenaires sociaux
L'évolution des coûts unitaires de main-d'œuvre, qui ont enregistré un taux de croissance légèrement négatif au cours du troisième trimestre de 1998, a également contribué à la diminution du taux d'inflation.

Etape 2 : Prévision d'un risque inflationniste pour 'responsabiliser' les partenaires sociaux
En ce qui concerne les pressions à la hausse [des prix], des augmentations de salaires […] pourraient représenter des risques inflationnistes pour l'avenir.

Etape 3 : Satisfecit final pour maintenir la confiance des investisseurs
Dans l'ensemble, cependant, les perspectives relatives à l'évolution des prix dans la zone euro peuvent être considérées comme étant globalement équilibrées.

4. Conclusion

Ainsi, on a pu observer comment l'argument macro-économique de stabilité monétaire est utilisé par la BCE pour décréter une « nécessaire modération salariale ». Elle part du présupposé que le chômage est lié à un problème de coût du travail trop élevé, alors même que cela n'a jamais été prouvé empiriquement. La politique macro-économique, dite de désinflation compétitive est entendue comme seule politique possible, tout comme l'est la théorie néoclassique au plan scientifique. Il serait alors intéressant de comparer plus en détails cette vision macro-monétariste de la BCE, indépendante qui a abandonné la variable « croissance » à celle du Conseil ECOFIN inscrite dans les GOPE sur le « cercle vertueux croissance/emploi », plus mitigée, puisque inscrite dans un processus décisionnel contrôlé. A ces deux variantes d'une politique de rigueur pourrait-il s'opposer plus radicalement une politique monétaire au service de l'intérêt général ?

On a également pu montrer comment la BCE va dans le sens d'un traitement du chômage par des politiques structurelles à des fins de stabilité macro-économique. L'argument du chômage « structurel » (dû aux imperfections du « marché du travail ») lui permet de ne pas concevoir de coordination des politiques macro-économiques. Pour le moment, ces arguments de contraintes macro-économiques sont autant de prétextes utilisés pour « responsabiliser », *i.e.* « modérer » les revendications des syndicats et de la « sphère sociale » dans son ensemble. Là encore, il serait pertinent pour des recherches à venir d'effectuer des comparaisons entre les « politiques structurelles » du marché du travail proposées par la BCE et celles prescrites par la Stratégie européenne pour l'emploi (SEE). Ceci apparaît d'autant plus pertinent du fait de la récente rationalisation des GOPE et des PANE en un processus parallèle. Quelle serait alors une politique publique d'emploi alternative au niveau européen ?

Plus généralement, face à cette doxa néo-libérale, on pourrait s'interroger sur ce que pourrait être une vision « sociale » où la Banque n'imposerait pas ses recommandations monétaristes, et où le progrès des droits sociaux ne seraient pas conçu comme subsidiaire mais bien comme conditions du bon fonctionnement de l'économie.

Annexe : Tableau des 40 premières formes du corpus BCE (1999-2002)

Fréquence	*Forme*	*Rang*
12437	%	1
9985	taux	2
9890	euro	3
8156	zone	4
5340	croissance	5
5005	prix	6
4398	bce	7
4105	terme	8
3921	mois	9
3863	données	10
3629	cours	11
3065	monétaire	12
3051	intérêt	13
2695	trimestre	14

Fréquence	*Forme*	*Rang*
2644	pays	15
2610	marché	16
2417	niveau	17
2403	opérations	18
2355	tableau	19
2322	évolution	20
2300	conseil	21
2231	hausse	22
2228	fin	23
2219	janvier	24
2204	secteur	25
2129	titres	26
2106	évolutions	27
2064	milliards	28

Fréquence	*Forme*	*Rang*
2024	statistiques	29
2002	période	30
1998	politique	31
1997	facilité	32
1979	euros	33
1979	eurosystème	34
1966	base	35
1923	économique	36
1827	refinancement	37
1824	long	38
1778	année	39
1771	marchés	40
1749	gouverneurs	41

Références

Artis M.J. (2003). EMU Four Years On. *mimeo*.

Artus P. et Wyplosz C. (2002). *La Banque centrale européenne*. Rapport du Conseil d'analyse économique. La Documentation française.

Barbier J.-C. et Nadel H. (2000). *La flexibilité du travail et de l'emploi*. Dominos. Flammarion.

BCE (2000). Monetary policy transmission in the Euro Area. *ECB Monthly Bulletin*. July 2000 : 43-58.

Buiter W. (1999). Alice in Euroland, *Journal of Common Market Studies*, vol. (37/2) : 181-209.

Bofinger P. (1999). La politique monétaire de la BCE au regard de l'article 105 du Traité, document de travail. PE 168 261, avril 1999.

Bourdieu P. (1997). L'architecte de l'euro. *Le Monde Diplomatique*, 27 février 1997.

Conseil de l'Union européenne (1999). Règlement CE n°530/1999 du Conseil du 9 mars 1999 relatif aux statistiques structurelles sur les salaires et les coûts de la main d'œuvre. JO L63 du 12 mars 1999 : 6-10.

Dufresne A. (1999). Rapport de la BCE - aspects sociaux. *mimeo*, décembre 1999.

Dufresne A. (2000a). Analyse des rapports annuels des banques centrales nationales et de la BCE concernant le marché du travail. *mimeo*, janvier 2000.

Dufresne A. (2000b). Une perspective bancaire en matière budgétaire. *mimeo*, août 2000.

Dufresne A. (2000c). La BCE. un acteur politique ? *L'année sociale 2000*. Institut de sociologie de l'ULB, juin 2000.

Dufresne A. (2001a). Les Grandes orientations de politiques (économiques?): un tournant institutionnel. Quels changements substantiels ?. *Revue belge de sécurité sociale*, vol. (3) : 597-620.

Dufresne A. (2001b). Quel bilan pour la Banque centrale européenne ? *Notabene,* vol. (119/2-6).

Dufresne A. (2001c). Les arguments de la BCE pour la privatisation et l'individualisation des retraites, *mimeo,* novembre 2001.

Dufresne A. (2002). Le rêve d'Oskar Lafontaine : une occasion pour la coordination des politiques économiques ?. In Degryse C. et Pochet P. (Eds), *Bilan social de l'Union européenne*. ISE-OSE-Saltsa : 85-113.

Gobin C. (2003). L'Union européenne : l'institution politique est évanescente. le syndicat est un partenaire. le travailleur un problème. où est passé l'acteur ? In *les pré-actes du colloque internatio-*

nal. La représentation de l'acteur au travail, organisé par le CLERSE (Univ. Lille 1). Villeneuve d'Ascq. les 20 et 21 novembre 2003. (Tome 1 : partie 3).

Hall P.A. et Franzese R.J. (1998). Mixed Signals: Central Bank Independence. Coordinated Wage-Bargaining and European Monetary Union . *International Organization*, vol (52/3) : 505-535.

Hetzel A.-M., Lefèvre J. *et al.* (1998). *Le syndicalisme à mots découverts*. Syllepse.

Hoang-Ngoc L. (1996). Salaires et emploi – une critique de la pensée unique. Syros.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Dunod.

Magnette P. (2000). Quis custodes custodiet? La Banque centrale europénne entre independance et responsabilité . *Cahiers européens de Sciences Po*., vol. (1). Centre Européen de Sciences Po.

Parlement européen (1999). *Résolution sur les conséquences constitutionnelles de l'UEM dans le contexte de l'élargissement de l'UEM du 11 février 1999*. doc. PE A4-0030/99.

Parlement européen et Conseil de l'Union européenne (2001). Proposition de règlement du Parlement européen et du Conseil relatif à l'indice du coût de la main d'œuvre. COM (2001) 418 final du 23 juillet 2001. JO C 304 E du 30 octobre 2001 : 0184-0187 (http://europa.eu.int/eur-lex/fr/com/pdf/2001/fr_501PC0418.pdf).

Raveaud G. (2003). "The European Employment Policy: from ends to means?", Working Paper, June, fothcoming in Salais R., Villeneuve R. eds., 2004, Europe and the politics of capabilities, Cambridge, Cambridge University Press.

Riche P. et Wyplosz C. (1993). *L'Union monétaire de l'Europe*. Le Seuil.

Salais R. (2003a). La politique des indicateurs. Du taux de chômage au taux d'emploi dans la stratégie européenne pour l'emploi (SEE). In Zimmermann (sous la direction de), *Actions publiques et sciences sociales*. Ed. de la Maison des Sciences de l'Homme. (A paraître).

Tietmeyer H. (1999). *Economie sociale de marché et stabilité monétaire*. Economica.

Choice of Text Analysis Software in Organization Research: Insight from a Multi-dimensional Scaling (MDS) Analysis

Vincent J. Duriau[1], Rhonda K. Reger[2]

[1,2]ITAM, Av. Camino a Santa Teresa, #930, Del. Magdalena Contreras, México, D. F. 10700
vduriau@itam.mx, vduriau@rhsmith.umd.edu
[2]R.H. Smith School of Business, University of Maryland, College Park, MD 20742-1815
rreger@rhsmith.umd.edu

Abstract

Content analytic approaches have made great strides in organization studies in the course of the past two decades. However, management scholars are often challenged to choose the best software solution to implement their research project. In order to assist in such a critical methodological decision, we develop a more thorough and comprehensive categorization of computer-aided text analysis (CATA) software.

First we summarize the strengths and limitations of CATA as it has been applied in organization studies. Second, we review typologies of such software that have been proposed. Finally, we report on the results of a multidimensional scaling (MDS) analysis of 33 CATA packages. These findings should help management researchers to make better informed decisions in their choice of CATA software.

Résumé

Les approches d'analyse de contenu ont fait d'importants progrès pour la recherche en gestion au cours des deux dernières décennies. Néanmoins, les chercheurs en management ont souvent des difficultés à choisir le meilleur logiciel pour leur projet. À cet effet, nous développons une catégorisation plus rigoureuse et complète des logiciels d'analyse de contenu.

Nous commençons par résumer les avantages et les limitations des logiciels d'analyse de contenu dans leur application à la recherche en gestion. Ensuite, nous résumons les typologies existantes de ces logiciels. Finalement, nous décrivons les résultats d'une analyse « multidimensional scaling » de 33 logiciels d'analyse de contenu. Ces résultats devraient assister les chercheurs en gestion à prendre de meilleures decisions dans leur choix de logiciels.

Keywords: computer-aided text analysis, multidimensional scaling.

1. Introduction

The computer revolution has contributed to the proliferation of content analytic methodologies in organization research (Weitzman and Miles, 1995; Kelle, 1995; Roberts, 1997; Tesch, 1991). New programs with enhanced capabilities have made subtle analysis of large amounts of quantitative and qualitative data possible (Gephardt, 1993; Lissack, 1998). We review the computer-aided text analysis (CATA) packages available today and use multidimensional scaling (MDS) to make sense of this vast array. With the breathtaking pace of new feature introductions, we use the MDS results to develop criteria that can help management researchers evaluate future software releases.

Our interest centered on two major questions: 1) How have CATA software been applied in organization studies? And , 2) How can CATA packages be categorized to help management researchers make a better informed decision to implement their projects?

2. Computer-aided Text Analysis in Management Research

2.1 *Definition of computer-aided text analysis*

Several definitions of computer-aided text analysis (CATA) have been proposed (see, e.g., Kabanoff, 1997; Mossholder *et al.*, 1995; Wolfe *et al.*, 1993). Since multiple methodologies and technologies have been included under the CATA rubric, we have adopted Wolfe *et al.*'s inclusive definition: CATA is constituted by software programs that "facilitate the analysis of textual data" (Wolfe *et al.*, 1993, 638). Most authors use the expression computer-aided textual analysis (CATA) because its short form is preferred to the unfortunate acronym for computer-aided content analysis. Nonetheless, the terms are interchangeable.

2.2. *Strengths of CATA*

One way to understand the contribution of computers to content analysis is to contrast computer-aided and manual approaches (Kelle, 1995). In addition to easier data manipulation, the use of software affords several analytical advantages that greatly enhance the methodology. First, computerization allows the manipulation of large data sets (Gephart, 1991; Lissack, 1998; Morris, 1994; Wolfe *et al.*, 1993). The complexity and potential interrelationships of concepts increase exponentially with the quantity of data. Software programs offer features for organizing, searching, retrieving, and linking data that renders the process of handling a large project much more manageable. For instance, Lissack (1998) described how parsing software can be used to sample concepts from a large corpus. This sampling approach allows the researcher to content analyze a reasonable amount of data representative of the initial corpus.

Second, computers reduce the time and cost of undertaking content analytic projects (Mossholder *et al.*, 1995). Time savings stem from the minimization of the coding task, the reduction in coder training, the elimination of inter-rater checks, and the ease of running multiple analyses (Carley, 1997).

Third, the use of computers addresses several of the reliability concerns associated with manual coding (Morris, 1994; Gephart and Wolfe, 1989; Wolfe *et al.*, 1993). Coding rules are made explicit which ensures perfect reliability and comparability of results across texts.

There are encouraging results that the semantic validity possible with manual coding using multiple coders can be achieved at a lower overall cost with CATA (Kabanoff, 1996; Kelle, 1995; Morris, 1994). Morris (1994) tested the validity and reliability of manual and computerized approaches. Using mission statement data from Pearce and David (1987), she compared the outcome of computerized coding in ZyIndex, a text management software, with that of a panel of six graduate business students. She found that results from ZyIndex and the human coders agreed at an acceptable level and that computerized coding yielded an acceptable level of semantic validity (Morris, 1994).

Finally, the use of network concepts have been one of the most exciting developments of the past few years in CATA research (Kelle, 1995). New linkage features between text, memos, and codes such as hyperlinks and graphical tools apply to the areas of theory building, hypothesis testing, and integration of qualitative and quantitative analysis. These developments of CATA seem particularly apt to quell concerns about the decontextualization of results that is inherent to a methodology based on coding and retrieval (Dey, 1995; Prein and

Kelle, 1995). Gephart (1993) also observed the methodological fit of computer-aided text analysis with grounded theorizing because information can be retrieved in meaningful ways that allows for the emergence of grounded hypotheses (see also, Wolfe *et al.*, 1993).

2.3. *Limitations of CATA*

There are still a number of questions associated with the use of CATA (Carley, 1997; Gephardt, 1991; Morris, 1994; Tallerico, 1991). The debate regarding manifest versus latent content comes first and foremost (Woodrum, 1984). Although CATA software is increasing in sophistication, measuring content with computers will miss some latent aspects within texts such as tone or irony of expression (Morris, 1994). However, human coders also exhibit low reliability for latent content. Further, we believe that the significance of these problems may be over-estimated for business texts appearing in corporate documents. These documents are usually written for clarity because they are read by people from around the world. Still, validity may be a particularly critical issue when dealing with metaphors, homonyms, colloquialisms (Carley, 1997), and other aspects of natural language (Morris, 1994).

There are three additional issues in implementing CATA. First, retrieval capabilities may be insufficient in certain software categories (Gephart and Wolfe, 1989). Second, researchers should avoid the pitfalls of the false sense of security afforded by a computer and the justification fallacy of using a computer program (Tallerico, 1991). Computerization will never replace human judgment in all cases. Finally, the fragmentation of the field of CATA software makes the choice of the appropriate package difficult (Kabanoff, 1996).

Categorization of CATA Software

Several interesting and important CATA typologies have been proposed. For instance, Tesch (1991 and 1990) introduced a typology based on the two dimensions of methodology and technology that has been adapted by several other authors (see, e.g., Wolfe *et al.*, 1993; Roberts, 1997). She made the – now blurring – distinction between commercial and academic software (Tesch, 1991 and 1990). Then, she proceeded with categorizing the academic packages.

More recently, Weitzman and Miles (1995) established a practical list of types of CATA packages, ranging from simple to more complex programs. The typologies of Tesch (1991) and Weitzman and Miles (1995) are summarized in Table 1.

Tesch (1991)

Type of software, main features	Examples*
Descriptive/interpretive analysis. The main goal is to discover the meaning of the phenomena (patterns, types). Data is kept unstructured and is flexibly manipulated. Main functions are coding and retrieval of text. Enhancements may include frequency count, chain, co-occurrence, advanced search, memoing, and quantitative option.	Ethnograph, Qualpro, Textbase Alpha, TAP, MARTIN, LTT Ethnoscript
Theory-building research. The objective is to elicit concepts/linkages in the data. Basic functions of text coding and retrieval are included. Key functions are concept identification, co-occurrence.	Kwalitan, HyperRESEARCH, NUDIST, AQUAD, ATLAS.Ti

Traditional content analysis or cultural analysis. Expansion of the traditional search function of word processors.	General Inquirer, TextPack, Word-Cruncher, FlexText
Linguistic Programs. Linguistic aspects of data.	CODEF, PLCA

* In 1991, these programs were running on MS-DOS or MacIntosh personal computers. Examples are given for the MS-DOS platform only.

Weitzmann and Miles (1995)

1. W*ord processors*, such as Word and WordPerfect, can be used for a variety of support tasks including handling field notes, transcribing interviews, preparing files, taking project notes (memoing), and writing reports.
2. *Text retrievers* specialize in finding and retrieving words, phrases, and characters strings. Some, such as ZyIndex, have content analytical features.
3. *Text base managers* are focused on the organization of textual data, including features such as sorting, searching, and retrieving. Asksam and WinMax are examples.
4. *Code and retrieve programs* are geared toward the coding and manipulation of text. Ethnograph is an example.
5. *Code-based theory builders*, such as NUD*IST, provide additional features including theory building tools, such as connections between codes, coding hierarchies, and formulation and testing of hypotheses.
6. *Conceptual network builders* focus on the network aspects of theory development and provide graphical features to support theoretical development

Table 1. CATA Typologies

Finally, Kabanoff (1997) suggested in his introduction to the *Journal of Organizational Behavior*'s special issue a typology of CATA-based management research using two dimensions: data sources and analytic methods. Data sources can be evoked – collected by the researcher in questionnaires or surveys, or natural – documents such as annual reports or newspaper articles. The second dimension ranges from quantitative to qualitative (Kabanoff, 1997 and 1996).

3. MDS Analysis of CATA Packages

Given the challenges facing management scholars to choose the best software solution to implement their content analytic project, we developed a more thorough and comprehensive categorization of existing computer-aided text analysis (CATA) software. We conducted an MDS analysis comparing the CATA packages referenced on the Georgia State University (GSU) web site on content analysis (http://www.gsu.edu/~wwwcom/content/csoftware/software_menu.html), using the dimensions from the three typologies discussed above as well as additional practical considerations.

3.1. *Methods*

It is difficult to provide a comprehensive list of all existing CATA software and the GSU web site was the most exhaustive reference that we could find. Among the 57 CATA packages described on the GSU web site, 24 were excluded from the MDS analysis because they did

not operate on the Windows operating system, were no longer supported, had a functional scope that was too broad, or did not have an English version.

We then established criteria on which to evaluate these packages. The choice of features was based on extant typologies (Kabanoff, 1997; Tesch, 1991; Weitzman and Miles, 1995) as well as a review of the documentation downloaded from CATA software suppliers' web sites. A comprehensive database was compiled using web sites' information, software demo versions, users' manuals, and contacts with software suppliers by e-mail. Due to its size, the complete database cannot be rendered in this paper, but an evaluation of dtSearch, TATOE, and NVivo according to 30 major features is shown on Table 2.

SOFTWARE PACKAGE	DTSEARCH	TATOE	NVIVO
Price	$199	Free	$425
Orders	1-800-483-4637	N/A	Online/phone
Technical support	(703) 413-3670	N/A	805-499-1325
Website	www.dtsearch.com	http://www.darmstadt.gmd.de/~rostek/tatoe.htm	www.qsr-software.com
Fax	(703) 413-3473	N/A	805 499 0871
E-mail	sales@dtsearch.com	alexa@zuma-manheim.de; rostek@darmasdt.gmb.de	nudist@sagepub.com
User group	N/A	N/A	YES
Demo version	YES	YES	YES
Data Input			
ASCII	YES	YES	YES
Productivity software	YES	N/A	N/A
HTML	YES	YES	N/A
Other	N/A	XML	RTF
Text manipulation			
Preparation	YES	YES	YES
Memoing/coding	N/A	YES	YES
Coding	N/A	YES	YES
Dictionaries	YES	YES	YES
Project management	N/A	YES	YES
Functionalities			

Searching	YES	YES	YES
Basic text statistics	YES	YES	YES
Content Analysis	N/A	YES	N/A
Qualitative Data Analysis	N/A	N/A	YES
Statistical Analysis			
Multiple regression	N/A	N/A	N/A
Cluster analysis	N/A	N/A	N/A
Multidimensional scaling	N/A	N/A	N/A
Other	N/A	N/A	N/A
Data output			
ASCII	N/A	N/A	YES
Productivity software	N/A	N/A	YES
HTML	N/A	YES	N/A
Statistical packages	N/A	YES	YES
Other	DBV, CSV, RTF	XML	RTF

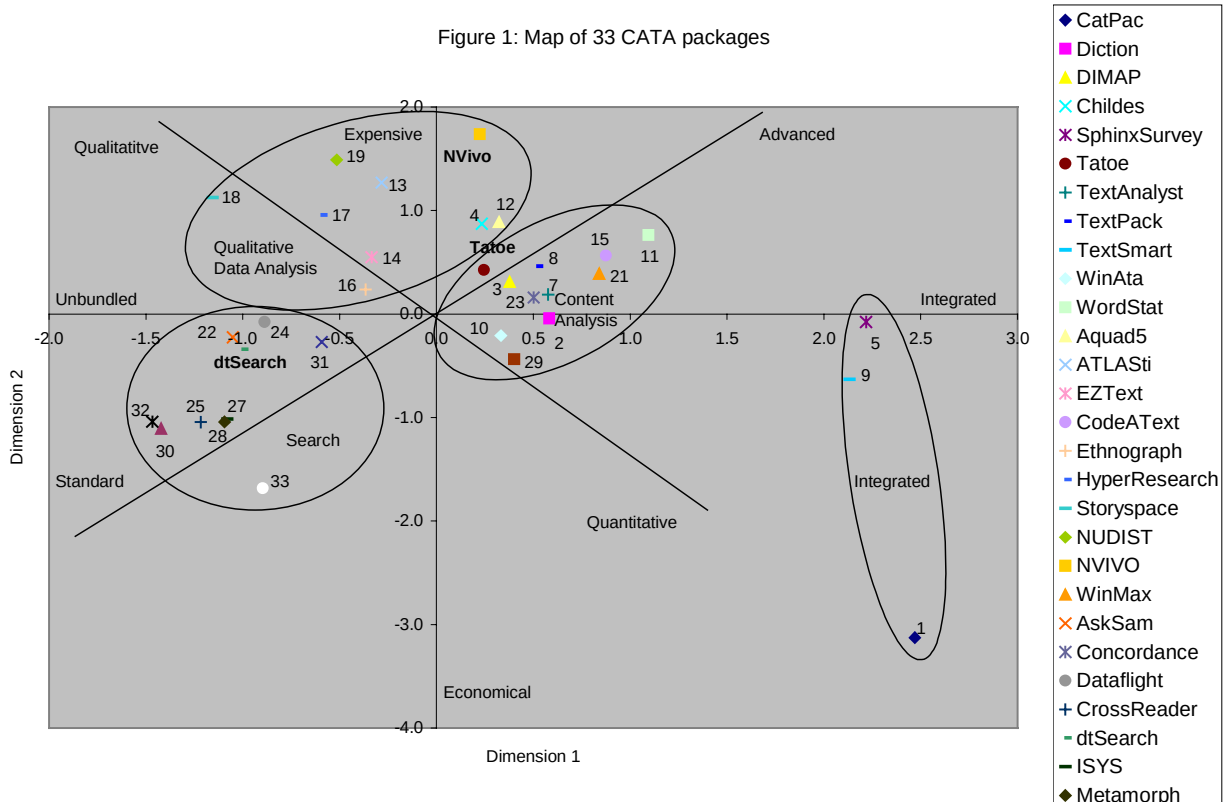
Table 2. Comparison of dtSearch, TATOE, and NVivo

To prepare the data for statistical processing, we indicated the presence of each feature in a software package by a binary count. To simplify the MDS analysis, features were then grouped in seven categories: price, technical support, data input, text management, content analytic features, statistical analysis, and data output. Consistent with extant typologies, we maintained the distinction between the functionalities of searching, basic text statistics, content analysis, and qualitative data analysis.

3.2. *Results*

The results of the MDS analysis confirmed and enriched the typologies previously discussed. First, the two-dimension analysis revealed several categorization dimensions: quantitative-qualitative, standard-advanced, unbundled-integrated, and economic-expensive (see Figure 1). Second, three clusters differentiating qualitative, quantitative, and search CATA packages also appeared, which were consistent with Tesch's (1991) and Weitzman and Miles' (1995) typologies. These results countered our expectations that the market for CATA packages was converging and that the segmentation valid a decade ago was blurring. In addition, CATPAC, SphinxSurvey, and TextSmart clustered into a new group of integrated content analysis packages. While the use of three dimensions improved the quality of the MDS analysis (stress=0.158 and R^2=0.901 for three dimensions, versus stress=0.254 and R^2=0.814 for two dimensions), we show the results for the two-dimension model, which provides a simpler taxonomy of CATA software packages for practical purposes.

Figure 1: Map of 33 CATA packages



4. Conclusion

We provide new insight for the evaluation and selection of CATA software. We identified and empirically tested the typologies proposed in the literature. The results of the multidimensional scaling (MDS) analysis largely confirmed the utility of previously proposed typologies, but also suggested important refinements. Additionally, the placement in two-dimensional space of 33 current and widely available CATA packages should benefit researchers in selecting the solution that is optimal for their project.

The advent of CATA software has led to significant benefits in terms of cost, flexibility, and reliability, while maintaining satisfactory levels of validity (Morris, 1994). In our opinion, the highest quality and most interesting content analytic research has used a combination of computer automation and human intervention to increase efficiency and perform more subtle and powerful analyses (see, e.g., Abrahamson and Park, 1994). We believe that this is the optimal approach to conduct sophisticated content analysis with the necessary level of scientific rigor. But, as with all research, the choice of methodology and design must be driven by the research purpose.

References

Abrahamson E. and Park C. (1994). Concealment of negative organizational outcome: An agency theory perspective. *Academy of Management Journal*, vol. (37): 1302-1334.

Carley K. (1997). Extracting team mental models through textual analysis. *Journal of Organizational Behavior*, vol. (18): 533-559.

Dey I. (1995). Reducing fragmentation in qualitative research. In Kelle U. (Ed.). *Computer Qualitative Data Analysis: Theory, Methods, and Practice*. Sage Publications: 69-79.

Gephart R.P. (1993). The textual approach: Risk and blame in disaster sensemaking. *Academy of Management Journal*, vol. (36): 1465-1514.

Gephardt R.P. (1991). Multiple approaches for tracking corporate social performance: Insights from a study of major industrial accidents. *Research in Corporate Social Performance and Policy*, vol. (12): 359-383.

Gephardt R.P. and Wolfe R.A. (1989). Qualitative data analysis: Three micro-supported approaches. In *Academy of Management Proceedings*: 382-386.

Kabanoff B. (1997). Introduction. Computers can read as well as count: Computer-aided text analysis in organizational research. *Journal of Organizational Behavior*, vol. (18): 507-511.

Kabanoff B. (1996). Computers can read as well as count: How computer-aided text analysis can benefit organizational research. *Trends in Organizational Behavior*, vol. (3): 1-21.

Kelle U. (1995). *Computer-aided qualitative data analysis: Theory, methods, and practices*. Sage Publications.

Lissack M.R. (1998). Concept sampling: A new twist for content analysis. *Organizational_Research Methods*, vol. (1): 484-504.

Morris R. (1994). Computerized content analysis in management research: A demonstration of advantages and limitations. *Journal of Management*, vol. (20): 903-931.

Mossholder K.W., Settoon R.P., Harris S.G. and Armenakis A.A. (1995). Measuring emotion in open-ended survey responses: An application of textual data analysis. *Journal of Management*, vol. (21): 335-355.

Pearce J.A. and David F. (1987). Corporate mission statements: The bottom line. *Academy of Management Executive*, vol. (1): 109-116.

Prein G. and Kelle U. (1995). Using linkages and networks for theory building. In Kelle U. (Ed.), *Computer qualitative data analysis: Theory, methods, and practice*: 69-79. Sage Publications.

Roberts C.W. (1997). *Text analysis for the social sciences: Methods for drawing statistical inferences from text and transcript*s. Lawrence Erlbaum Associates.

Tallerico M. (1991). Applications of qualitative analysis software: A view from the field. *Qualitative Sociology*, vol. (14): 275-285.

Tesch R. (1991). Introduction. *Qualitative Sociology*, vol. (14): 225-243.

Tesch R. (1990). *Qualitative research: Analysis types and software tools*. The Falmer Press.

Weitzman E.A. and Miles M.B. (1995). *Computer programs for qualitative data analysis*. Sage Publications: Thousand Oaks.

Wolfe R.A., Gephart R.P. and Johnson T.E. (1993). Computer-facilitated qualitative data analysis: Potential contributions to management research. *Journal of Management*, vol. (19): 637-660.

Woodrum E. (1984). "Mainstreaming" content analysis in the social science: Methodological advantages, obstacles, and solutions. *Social Science Research*, vol. (13): 1-19.

Corpus issus du Web : analyse des pertinences thématique et informationnelle

Louisette Emirkanian[1], Christophe Fouqueré[2], Fabrice Issac[3]

[1]Département de linguistique et de didactique des langues – UQAM – CP 8888 succursale 'Centre-ville'
Montréal (QC) H3C 3P8 – Canada – emirkanian.louisette@uqam.ca
[2]LIPN, CNRS UMR 7030 – Université Paris 13 – Villetaneuse – France
[3]LLI, CNRS UMR 7546 – Université Paris 13 – Villetaneuse – France

Abstract

The purpose of our study is to identify useful reformulation strategies for queries used in information retrieval. As part of an on-going research project, only one aspect of which is discussed here, we conducted a detailed study of five basic queries and of their variants obtained by morphological and synonymic enrichment. Although in general the use of synonymic and morphological variants improves recall, we have found that specifying the syntactic link between the terms in the query improves precision. We first describe the methodology used to assemble the corpus and comment on the data collected. We then examine the syntactic links between query terms and their correlation with thematic and informational relevance.

Résumé

Notre étude a pour objectif principal de mettre à jour des pistes pour la spécification de mécanismes de reformulation de requêtes facilitant la recherche d'information. Dans un projet de recherche dont nous présenterons ici un aspect, nous avons étudié de façon détaillée les résultats de cinq requêtes de base et de variantes obtenues par enrichissement morphologique et synonymique. Alors qu'en général la prise en compte de variantes synonymiques ou morphologiques permet d'améliorer le rappel, nous tenterons de montrer que la spécification du lien syntaxique entre les termes apporte une plus grande précision. Après avoir décrit la méthodologie de constitution de corpus issus du Web et commenté les données recueillies, nous examinerons les liens syntaxiques entre les termes des requêtes et leur rapport avec les pertinences thématique et informationnelle.

Mots-clés : recherche d'information, Web, pertinence thématique et informationnelle, syntaxe

1. Introduction

Nous sommes confrontés tous les jours à une masse importante d'informations. Le problème majeur qui se pose alors est d'être capable de repérer, dans cette masse d'informations, celle qui nous sera utile, celle qui répondra le plus précisément possible à notre besoin d'information. Le repérage d'information (RI) sur le Web présente des défis particuliers, en raison de la grande variété de domaines, genres et styles des documents. Les résultats d'une recherche sont souvent très nombreux et peu pertinents, dans le cas de requêtes constituées d'une courte liste de mots isolés, ou à l'inverse, trop peu nombreux dans le cas où l'on spécifie une requête par une coordination d'un trop grand nombre de termes, ou par l'utilisation d'expressions trop précises.

Jusqu'à récemment, les recherches dans le domaine du RI et celles dans le domaine du traitement automatique des langues n'entretenaient que peu de liens, chaque domaine développant

des outils spécifiques. Depuis peu, la convergence des deux domaines s'est faite et elle a porté ses fruits dans chacun d'eux (Jacquemin, 2000 ; Jacquemin et Zweigenbaum, 2000 ; Gaussier *et al.*, 2000 ; Strzalkowski, 1995 et 1999 ; Spärck Jones, 1999 ; Woods *et al.*, 2000). Des travaux récents tendent à prouver que l'utilisation de techniques de TAL en morphologie, syntaxe et sémantique permet d'améliorer la performance des systèmes de RI tant au niveau de l'indexation qu'à celui du repérage.

Ces connaissances linguistiques sont utilisées pour le découpage en unités linguistiques, pour l'étiquetage, l'analyse en constituants (et l'indexation de syntagmes), la reconnaissance de termes complexes, la désambiguïsation en contexte et l'extension et la reformulation de requêtes (Bouillon et *al.*, 2000).

Notre étude a pour objectif principal de mettre à jour des pistes pour la spécification de mécanismes de reformulation de requêtes facilitant la recherche d'information.

Dans un projet de recherche[1] dont nous présenterons ici un aspect, nous avons étudié de façon détaillée les résultats de cinq requêtes de base et de variantes obtenues par enrichissement morphologique et synonymique. Nous avons également porté une attention particulière à la dimension syntagmatique ; cette partie du projet fera l'objet de la présentation proposée. Alors qu'en général la prise en compte de variantes synonymiques ou morphologiques permet d'améliorer le rappel, nous tenterons de montrer que la spécification du lien syntaxique entre les termes apporte une plus grande précision.

Nous nous attacherons d'abord à décrire la méthodologie de constitution de corpus que nous avons utilisée en mettant l'accent sur les spécificités du Web ; nous commenterons ensuite les données recueillies pour nos cinq jeux de requêtes, puis nous examinerons les liens syntaxiques entre les termes et leur rapport avec la pertinence thématique et informationnelle.

2. Constitution du corpus & méthodologie

À partir de 5 besoins en information de type X Y, VOYAGE TIBET (où VOYAGE constitue X et TIBET, Y), « FUITE DES CERVEAUX », ÉTATS UNIS, VOL LUNE, MISSION ESPACE, PROMENADE PARIS, nous avons constitué 5 corpus de pages html extraites du Web (désormais corpus FUITE, TIBET, LUNE, ESPACE et PARIS). Des requêtes de la forme « X (NEAR préposition) NEAR Y » ont été exécutées pour constituer chacun des corpus. Ces requêtes ont été construites par variation à partir du besoin initial. Nous avons recherché les pages du Web contenant les mots de base dans l'ordre initial, dans un voisinage, avec ou sans préposition, avec des variations morphologiques et sémantiques sur X et Y ainsi que des variations sur la préposition. 31 prépositions ont été pour cela utilisées (la totalité), de même toutes les variations morphologiques ont été essayées, enfin nous avons identifié tous les synonymes des mots pleins ; par exemple dans le cas de la requête VOYAGE TIBET, *voyage* est remplacé par *voyages*, *séjour*, *voyager*, *trek*, etc. et *Tibet* par *tibétain*, *tibétaines*, etc. Dans le cas où le Y est un adjectif, la préposition pourra être présente ou non (*promenade parisienne* ou *promenade dans les arrondissements parisiens*). Nous avons un total de 14840 requêtes.

2.1. *Principe*

À partir d'une ou plusieurs requêtes, exprimées dans un langage propre à un moteur de recherche, un ensemble de pages est récupéré puis stocké et transcodé. L'outil est en fait

[1] Ce projet a été financé par la coopération franco-québécoise (Ministère des Relations Internationales du Québec et Ministère des Affaires Étrangères de la France).

constitué d'un ensemble modulaire de sous-programmes écrits en langage PERL.

Un premier sous-programme permet de générer une liste de requêtes pour un moteur spécifique à partir de fichiers de configurations. Il effectue toutes les combinaisons possibles entre les différents éléments (mots-clés ou variantes de ceux-ci) de la requête et y associe un ou plusieurs connecteurs spécifiant en particulier la proximité des mots-clés dans la page recherchée. Un deuxième composant interroge alors un moteur de recherche et récupère toutes les adresses des pages correspondant aux différentes requêtes. Toutefois, la liste de pages réellement récupérables n'est pas exhaustive : ainsi, à partir d'une requête sur le moteur de recherche Altavista il n'est possible de récupérer qu'au plus 1010 URL. Les adresses apparaissant en double sont ensuite éliminées. Enfin, un dernier composant récupère la page elle-même (quand celle-ci est effectivement disponible !) et transcode le résultat dans un format XML apte à intégrer des informations supplémentaires (adresse de la page, date de récupération, requêtes associées, etc.). Hormis les quelques cas de fichiers de style (type css), il ne s'agit que de pages au format initial html. L'ensemble de ces pages constitue une photo instantanée des pages récupérables via un moteur de recherche spécifique (en l'occurrence AltaVista).

2.2. *Choix du moteur*

Le module essentiel au système a pour tâche d'analyser le résultat d'une requête d'un moteur de recherche. Ce résultat est une ou plus généralement plusieurs pages Web. Nous avons choisi pour cela un mécanisme capable tout à la fois d'émettre une requête, de récupérer les pages résultats, et d'extraire de ces pages les URL. L'outil est assez souple pour pouvoir être adapté à de nombreux moteurs. Nous avons choisi d'utiliser Altavista plutôt que Google, qui offre actuellement sans doute le meilleur classement et une meilleure couverture, car ce moteur permet de gérer plus aisément la notion de proximité entre mots clés. Ainsi l'opérateur « NEAR » ajoute à l'opérateur « AND » la contrainte d'une distance maximale de 10 mots entre deux mots-clés d'une requête.

2.3. *Corpus résultant*

Le résultat de l'aspiration des corpus est résumé dans le tableau ci-dessous, où nous avons précisé pour chaque corpus créé le nombre total de mots, la taille (en mégaoctets) du corpus, et enfin le nombre de pages de sites Web qui auront pu être analysées. Les deux dernières colonnes indiquent le nombre moyen de mots par page et son écart type. Il y a lieu de noter la variabilité très importante de ces données selon le thème initial. Cette variabilité n'est pas due à une limitation de nos outils mais bien à la variabilité intrinsèque du Web (pages journalistiques ou économiques, pages à caractère touristique, pages personnelles, etc.).

La variabilité en termes de nombre de pages reflète l'importance d'utiliser le Web comme véhicule de l'information liée à ce thème. Ainsi le thème PARIS renvoie à de très nombreuses pages à caractère journalistique, le thème ESPACE à une très forte volonté de diffusion électronique de l'information scientifique et technique. Si le nombre moyen de mots par page est sensiblement égal, cela résulte du désir des concepteurs de sites Web de découper l'information en pages de taille raisonnable tant relativement à la lecture qu'afin d'éviter un délai de chargement trop important. Toutefois, l'écart type montre que l'affirmation précédente laisse place à de très nombreuses exceptions.

CORPUS	TAILLE (MO)	Nb mots	Nb de pages	Nb de mots / page	Écart Type du nb de mots / page
FUITE	17	908 032	187	3 883	6 705
ESPACE	407	24 905 110	8 291	2 042	6 288
LUNE	187	10 978 288	3 522	2 444	9 186
TIBET	70	3 829 657	1 507	1 779	3 751
PARIS	1 058	60 588 606	20 017	2 364	8 732
TOTAL	***1 739***	***101 209 693***	***33 523***	***3 019***	

Tableau 1. Caractéristique des corpus

2.4. *Post traitement*

Un post traitement est effectué afin de repérer dans le texte :

1. les phrases contenant une séquence correspondant à une des requêtes ayant permis de récupérer la page,
2. les phrases contenant une séquence correspondant à une des requêtes alors que celle-ci n'avait pas permis de récupérer la page,
3. les mots de la requête n'appartenant pas à une séquence.

Pour ce faire la page HTML est transcodée en XML. La structure permet de repérer, outre les informations contenues dans le résumé, les phrases et la liste des requêtes dont le motif syntaxique est inclus dans la phrase. Pour être repérée, une phrase doit avoir en son sein les différents éléments d'une requête dans l'ordre initial. Il est à noter que le découpage en phrases est un exercice non trivial sur ce type de données, par conséquent nous donnons à la notion de phrase un sens relativement lâche. Toutefois, cette étape de découpage en phrases est indispensable dès lors que nous souhaitons prendre en compte la préservation de la structure syntaxique présente initialement dans la requête.

3. Analyse

Afin d'analyser le corpus obtenu du point de vue de la pertinence thématique et informationnelle, nous avons élaboré une interface graphique permettant de choisir son corpus, de le parcourir puis de l'annoter.

3.1. *Pertinence informationnelle*

La *pertinence informationnelle* est la pertinence classique du RI, qui identifie les documents répondant complètement ou partiellement au besoin de l'utilisateur, sans nécessairement établir de liens avec la formulation particulière de la requête. Étaient jugés pertinents au niveau informationnel, pour le corpus FUITE, les documents traitant de la fuite des cerveaux vers les États-Unis ; pour le corpus TIBET, tous les documents qui donnaient des informations pratiques pour l'organisation d'un voyage touristique au Tibet ; pour le corpus ESPACE, tous les documents donnant des informations scientifiques sur les missions dans l'espace intersidéral ; pour le corpus LUNE, tous les documents donnant des informations scientifiques ou des détails sur les vols (effectués ou à venir) vers la lune ; enfin pour le corpus PARIS, tous les documents donnant des informations pratiques permettant d'organiser des promenades dans Paris. Nous avons retenu trois niveaux de pertinence, la seule justification de cette extrême limitation est de nous permettre d'avoir des outils qui nous indiqueront les grandes tendances des résultats :

- totalement pertinent (**TP**) : la page concerne en totalité le sujet concerné par la requête. On peut y parler ainsi de vol spatial (corpus ESPACE), de balade en bateaux-mouches dans Paris (corpus PARIS) ;
- partiellement pertinent (**PP**) : la page contient au moins un paragraphe répondant à la requête. C'est le cas par exemple dans les sites d'agences de voyages (corpus TIBET), quand, à coté de safaris kenyans, on mentionne des trekkings au Tibet ;
- non pertinent **(NP)** : la page ne répond pas à la requête. Cela ne signifie pas que les mots constituant la requête (voire le schéma associé) ne soient pas présents dans la page. Il peut s'agir d'un emploi figuratif du thème de la requête.

Pour en avoir fait l'expérience, nous savons que deux facteurs déterminent la "réussite" d'une requête : la pertinence de la page récupérée relativement à l'objectif initial de la requête, la pertinence de l'ordre dans lequel les pages récupérées nous sont exposées.

Le premier de ces facteurs est fondamentalement subjectif : des pages *a priori* très différentes sur le fond peuvent paraître pertinentes pour l'utilisateur qui aura effectué sa recherche. Il en est ainsi par exemple pour le corpus ESPACE, s'agit-il d'aéronautique ? de littérature ? d'ésotérisme ? d'espace architectural ? urbain ? Nous avons pu retrouver toutes ces thématiques lors de l'analyse manuelle que nous avons faite du corpus. Dans cette étude, il n'était pas envisageable de limiter la thématique dès lors que cette limitation n'apparaissait pas dans la spécification de la requête. Toutefois, c'est une optique qu'il est important d'envisager dans cette situation. Pour en revenir à notre projet, nous avons volontairement simplifié l'étude du contenu informationnel des pages.

À partir de cette analyse informationnelle, et des rangs associés à la page par le moteur de recherche, il nous était dès lors possible d'évaluer précisément les divers aspects entrant en jeu lors de la reformulation de requêtes. Il s'agit en effet d'évaluer le rappel, c'est à dire la proportion de pages pertinentes récupérées par chaque requête, et la précision, c'est à dire la proportion de pages pertinentes relativement à l'ensemble des pages récupérables. Plus précisément nous avons cherché à évaluer l'impact de la correction syntaxique de la requête ainsi que celui lié à la détermination de la préposition dans ce cadre. Dans un deuxième temps, nous avons intégré à ce calcul de pertinence le rang associé à la page en cherchant à savoir dans quelle mesure les algorithmes utilisés dans les moteurs de recherche correspondaient à nos premières évaluations. Il va de soi que ces informations sont largement dépendantes du moteur de recherche choisi pour l'expérience. Ainsi la technologie de ramassage, les calculs d'indexation du moteur et les critères utilisés par celui-ci pour le classement des réponses constituent-ils autant de biais relativement à notre expérience. Toutefois, nous considérons que le nombre de pages retenues (1010 au maximum) relativise ce biais.

Les requêtes syntaxiquement incorrectes donnent des résultats non pertinents : le rappel est très faible. C'est notoirement le cas pour le corpus FUITE DES CERVEAUX qui, avec 8462 requêtes pour seulement 182 pages distinctes récupérables, contient un nombre impressionnant de requêtes syntaxiquement incorrectes ne permettant de ne récupérer aucune page. De fait, 95 requêtes seulement sont de ce point de vue "utiles", chaque page étant récupérable en moyenne par 2 requêtes. Dans ce corpus, les prépositions syntaxiquement invalides (*chez les États-Unis*) sont des échecs du point de vue de la récupération. Toutefois, parce que le schéma associé n'est pas toujours respecté dans la phrase réelle, certaines requêtes syntaxiquement incorrectes permettent de récupérer des pages pertinentes non récupérées par d'autres requêtes. Ces cas sont suffisamment rares pour que la sélection de requêtes sur le critère de plausibilité syntaxique soit justifiée. Le cas des requêtes sans préposition reste particulier : le rappel y est (naturellement) très important, mais la précision est assez faible.

Enfin, la grande majorité des pages récupérées le sont aussi par le biais de requêtes avec prépositions. Les requêtes avec prépositions syntaxiquement pertinentes donnent des résultats au moins partiellement pertinents. Dans le cas du corpus ESPACE par exemple, seules 58 des 98 requêtes permettent la récupération de pages.

Enfin, nous avons cherché à analyser la "vitesse de couverture" des pages totalement ou partiellement pertinentes à partir du « facteur de qualité » associé aux requêtes. En associant degré de pertinence et rang, nous pouvons en effet calculer un facteur de qualité associé à la requête. Ce facteur est égal, pour une requête donnée, à la somme des pertinences pondérées par le logarithme du rang divisé par la somme des logarithmes des rangs. L'utilisation du logarithme permettant de minimiser l'effet du rang par rapport à la pertinence. Un résultat se rapprochant de 1 montre une adéquation entre le classement du moteur de recherche et notre propre classement. Pour cela, nous avons réanalysé l'ensemble des résultats en tenant compte du facteur de qualité. Afin de comprendre dans quelle mesure les requêtes pertinentes permettaient d'obtenir les résultats, nous avons cherché à couvrir l'ensemble des pages à partir des pages accessibles des requêtes en commençant par les requêtes de degré de qualité le plus élevé. Il est apparu clairement, pour les cinq thèmes étudiés, que 5 % des requêtes suffisent à récupérer 50 % des pages totalement ou partiellement pertinentes. Cela justifie indirectement la technique de la reformulation de requête. En effet, sous réserve de définir judicieusement ces 5 % de requêtes, la reformulation nous permet à peu de frais, i.e. modulo l'analyse par le moteur de recherche de ces requêtes supplémentaires, d'obtenir une liste significative (pour l'utilisateur) de pages jugées pertinentes. Qui plus est, la reformulation avec précision syntaxique (p.e. ajout de la préposition) non seulement permettra un rappel significatif, mais encore aura une précision correcte (eu égard à la masse de documents contenus dans le Web).

3.2. *Pertinence thématique locale*

La pertinence thématique locale est quant à elle rattachée à un passage contenant les termes de la requête, et non au document dans son ensemble, et identifie si ces termes sont employés dans un sens compatible avec le besoin en information et s'ils constituent l'objet du discours rattaché au passage. Par exemple, pour le corpus TIBET, *voyage au Tibet* sera jugé thématiquement pertinent au même titre que *voyage en terre tibétaine*, ou encore *voyage qui permet de découvrir le Tibet* ; en revanche, *marche pour le Tibet* sera jugé non pertinent.

Nous avons isolé un sous-ensemble représentatif pour chacun des 5 besoins en information. Nous n'avons retenu que les occurrences dans lesquelles les éléments X (préposition) Y apparaissent dans l'ordre de la requête et au sein de la même phrase. Nous avons ainsi obtenu 3174 occurrences réparties dans 1873 pages différentes**.** En effet, nous trouvons fréquemment des documents contenant plusieurs occurrences des termes de la requête. Ces occurrences se répartissent comme suit dans les différents corpus.

	Nombre de pages	**Nombre d'occurrences**	**Proportion d'occurrences pertinentes au niveau thématique**	**Répartition des occurrences au niveau informationnel**		
				TP	**PP**	**NP**
FUITE	92	116	79 %	23 %	56 %	21 %
TIBET	583	1 051	62 %	46 %	13 %	41 %
LUNE	492	776	70 %	31 %	15 %	54 %
ESPACE	282	540	68 %	34 %	38 %	28 %
PARIS	424	691	43 %	9 %	23 %	68 %
TOTAL	***1 873***	***3 174***	***62 %***	***31 %***	***22 %***	***47 %***

Tableau 2. Évaluation des pertinences informationnelle et thématique

Si l'on compare les proportions d'occurrences pertinentes au niveau thématique à celles des occurrences totalement ou partiellement pertinentes au niveau informationnel, on observe des similitudes sauf dans le cas du corpus LUNE. En effet, dans ce corpus, de nombreuses occurrences du type '*voyage dans la lune, vers la lune, sur la lune*' étaient correctes au niveau thématique mais le document dans lequel elles se trouvaient traitait d'œuvres littéraires ou musicales.

Il est également intéressant d'évaluer la correspondance entre d'une part les occurrences pertinentes au niveau thématique et d'autre part les occurrences totalement pertinentes et partiellement pertinentes au niveau informationnel, en d'autres mots, de vérifier s'il y a correspondance entre pertinence thématique et informationnelle**.** Sur l'ensemble du corpus, sur les 1967 occurrences pertinentes au niveau thématique, 72 % sont totalement ou partiellement pertinentes au niveau informationnel ; la proportion est inversée pour les séquences non pertinentes au niveau thématique ; en effet, seulement 20 % d'entre elles sont totalement ou partiellement pertinentes au niveau informationnel.

3.3. *Liens syntaxiques et pertinences informationnelle et thématique*

Nous avons réparti les occurrences obtenues en six catégories de liens syntaxiques ; elles sont liées au fait d'une part que nous autorisons différentes catégories pour X (noms, verbes) ou Y (noms, adjectifs) et d'autre part que nos requêtes contiennent l'élément NEAR.

- *N1-SP* :

  Y est dans un syntagme prépositionnel directement rattaché à X (*fuite des cerveaux vers les États-Unis, balade dans Paris, mission dans l'espace, vol vers la lune*, etc.)

- *N1-adj* :

  Le terme Y est un adjectif qui modifie directement X (*voyage tibétain, mission lunaire, spatiale, balade parisienne*, etc.)

- *P-N2* :

  Le terme Y est dans un syntagme prépositionnel non relié à X (*Au cours du vol de Gemini 4, Edward White effectua lui aussi une sortie dans l'espace, le 3 juin 1965.)* (http://www.multimania.com/msegret/laconqu.htm)

- *AUCUN LIEN* :

  Les trois termes X, préposition, Y ne sont pas reliés syntaxiquement. (*Un voyage très complet qui permet de découvrir à la fois le Tibet central et l'ancien royaume de Gugé*.) (http://tirawa.com/voyages/himalaya/tibet/tibet-703-lhassa-mont-kailash/index.html)

- *V-SP :*

  Le terme X est un verbe ; le terme Y est dans un syntagme prépositionnel rattaché au verbe (*Voyager au Tibet*).

- *V-SN :*

  Le terme X est un verbe ; Y est dans un syntagme nominal argument du verbe (*Ils ont parcouru le Tibet*).

Le tableau suivant présente une répartition des occurrences selon le schéma syntaxique dans lequel les éléments de la requête apparaissent.

Structures	Nombre total d'occurrences	Proportion d'occurrences thématiquement correctes	Répartition des occurrences au niveau informationnel		
			TP	MP	NP
N1-SP	1 128	77 %	38 %	22 %	40 %
N1-ADJ	280	98 %	35 %	31 %	34 %
V-SN	40	95 %	13 %	37 %	50 %
V-SP	213	79 %	22 %	25 %	53 %
P-N2	890	46 %	30 %	18 %	52 %
Aucun lien	623	33 %	24 %	19 %	57 %
TOTAL	***3 174***	***62 %***	***31 %***	***22 %***	***47 %***

Tableau 3. Évaluation des pertinences thématique et informationnelle selon les structures

Si l'on regroupe les cas où les éléments X et Y sont liés syntaxiquement (i.e. N1-SP, N1-ADJ, V-SN, V-SP) et qu'on les oppose au cas P-N2 (le syntagme prépositionnel n'est pas rattaché au N) et au cas où il n'y a aucun lien entre les termes de la requête, on constate que la proportion d'occurrences pertinentes au niveau thématique est plus élevée lorsque les éléments de la requête ont un lien syntaxique. Pour la pertinence informationnelle, les différences sont moins marquées. De plus, on note que lorsque l'élément X est un verbe, la proportion d'occurrences très pertinentes ou partiellement pertinentes au niveau informationnel est faible.

Nous avons montré, par ailleurs (Emirkanian et Chieze, à paraître), que si l'emploi de dérivés verbaux (pour l'élément X) et adjectivaux (pour l'élément Y) améliorait grandement le rappel, cela se faisait en général au détriment de la précision. Si nous laissons de côté les dérivés verbaux pour X et les dérivés adjectivaux pour Y, et que nous n'examinons que les cas où les termes X et Y sont des noms (voyage(s), trek(king), vol(s), mission(s), promenade(s), balade(s), Tibet, Paris, espace, etc.), nous obtenons les résultats suivants pour les trois schémas 'X préposition Y', 'X … préposition Y' et 'X … préposition … Y'. On remarque que la spécification du lien syntaxique entre les termes de la requête augmente à la fois la pertinence thématique et la pertinence informationnelle.

Structures	Nombre total d'occurrences	Proportion d'occurrences thématiquement correctes	Proportion d'occurrences très pertinentes au niveau informationnel
N1-SP	995	72 %	37 %
P-N2	590	45 %	30 %
Aucun lien	284	30 %	20 %

Tableau 4. Importance du lien syntaxique

4. Conclusion

L'analyse que nous avons menée a été effectuée en parallèle sur 5 corpus de pages extraites du Web. L'ampleur autant que la variabilité des corpus obtenus nous a permis d'effectuer à la fois des analyses automatiques (relatives à la précision et à la couverture de chaque type de requête) et des analyses manuelles (sur la pertinence thématique et informationnelle des pages extraites, sur les structures syntaxiques). Nous avons montré dans cet article les liens qui pouvaient exister entre pertinence thématique et informationnelle. Enfin, les résultats de cette étude indiquent que la spécification du lien syntaxique entre les termes de la requête permet

d'accroître la pertinence au niveau thématique. Il est important de noter que nous retrouvons cette augmentation au niveau de la pertinence informationnelle. La spécification du lien syntaxique par une préposition constitue dès lors l'une des pistes possibles pour la reformulation de requêtes dans l'objectif d'améliorer sensiblement la précision des réponses.

Références

Bouillon P., Fabre C., Sébillot P. et Jacqmin L. (2000). Apprentissage de ressources lexicales pour l'extension de requêtes. *TAL*, vol. (41/2) : 367-393.

Emirkanian L. et Chieze E. (à paraître) Variations morphologiques, syntaxiques, sémantiques et Repérage d'Information sur le Web. *Revue Québécoise de Linguistique*.

Gaussier E., Grefenstette G., Hull D. et Roux Cl. (2000). Recherche d'information en français et traitement automatique des langues. *TAL*, vol. (41/2) : 473-493.

Jacquemin Chr. (éd.) (2000). *Traitement automatique des langues pour la recherche d'information*. *TAL*, vol. (41/2).

Jacquemin Chr. et Zweigenbaum P. (2000). Traitement automatique des langues pour l'accès au contenu des documents. In Le Maître J., Charlet J. et Garbay C. (Eds), *Le document en sciences du traitement de l'information.* Cepadues : 71-109.

Spärck Jones K. (1999) The role of NLP in text retrieval. In Strzalkowski T., *Natural Language Information retrieval*. Kluwer : 1-24.

Strzalkowski T. (1995). Natural language information retrieval. *Information Processing & Management*, vol. (31/3) : 397-417.

Strzalkowski T. (Ed.) (1999). *Natural Language Information Retrieval*. Kluwer.

Woods W.A., Bookman L.A., Houston A., Kuhns R.J., Martin P. et Green S. (2000). Linguistic Knowledge can improve information retrieval. In *Proceedings of the 6th Applied Natural Language Processing Conference*.

Référentiels terminologiques adaptables au contexte : L'exemple d'un système de recherche d'informations dans une grande entreprise

Frédéric Erlos

Département intranet et organisation de Crédit Agricole S.A.
Doctorant en Sciences du Langage, Université de la Sorbonne nouvelle – Paris III
e-mail : frederic.erlos@credit-agricole-sa.fr

Abstract

The development of intranets in large enterprises has noticeably increased the constraints on information retrieval. The terminological knowledge bases, when calibrated for specific searching, make clear linguistic workings which generally elude intranet users. A real case enables us to present the main quantitative analysis of textual data used to the collection of words and semantic relations from corpus of institutional texts.

Résumé

Le développement des intranets dans les grandes entreprises a sensiblement accru les contraintes pesant sur la recherche d'informations. Les référentiels terminologiques, lorsqu'ils sont calibrés pour des processus de recherche précis, rendent explicites des fonctionnements langagiers qui échappent en général aux intranautes. Un cas concret permet de présenter les principales techniques lexicométriques utilisées pour la collecte d'unités et de relations sémantiques à partir de corpus de textes institutionnels.

Mots-clés : recherche d'informations, intranet, réseau sémantique, référentiel terminologique, lexicométrie, lexicographie, terminographie.

1. Introduction

Lorsque l'on envisage la question de la recherche d'informations sur un intranet[1], il devient urgent de répondre à des demandes telles que celles-ci : comment organiser l'information diffusée par telle unité en direction d'un public interne élargi qui ne possède pas nécessairement les mêmes références ? Quels mots retenir pour réaliser une expansion de requête efficace sur tel sujet ? Quels mots retenir pour constituer une liste fermée servant à qualifier et à diffuser l'information sur l'ensemble de l'intranet ?[2] En effet, la diffusion massive d'informations internes[3], autrefois cloisonnées, et la banalisation du traitement de l'information, qui n'est plus réservé à un groupe de spécialistes, obligent à trouver des réponses novatrices. Pour cela, il faut se donner les moyens de rompre avec le paradigme encore dominant en *Infor-*

[1] Pour définir ce que nous entendons par intranet, nous reprendrons la définition proposée par Esther Amar : « *Type de réseau utilisant les mêmes technologies qu'Internet (protocoles et applications TCP/IP), mais uniquement pour communiquer à l'intérieur d'une organisation sur son réseau local ou sur un grand réseau privé.* (…) » (Amar, 1997).

[2] On reconnaîtra ici les trois principaux moyens d'accéder à l'information sur un intranet : navigation par hyperliens, utilisation d'un moteur de recherche (mode *pull*) et information poussée vers l'utilisateur en fonction de ses centres d'intérêt (mode *push*).

[3] Il peut s'agir de plusieurs dizaines de sites particuliers comportant plusieurs milliers de pages, majoritairement sous formes de fichiers mis en ligne.

mation Retrieval d'un langage dans lequel les mots correspondent directement aux choses. Il s'agit en somme d'appréhender la recherche d'informations comme n'importe quelle situation de communication dotée de caractéristiques propres qui, si elles sont ignorées, nuisent à la qualité des réponses.

La voie que nous explorons suppose la mise en place de ressources linguistiques spécifiquement adaptées au cadre précis dans lequel se déroule le processus de recherche d'informations. Nous nommons ce type d'outils « référentiel terminologique dédié à la recherche d'informations » (RTRI pour la suite)[4]. Leur conception suppose de se doter d'un cadre théorique explicite dont la teneur sera exposée brièvement. Une série de trois exemples permettra ensuite de montrer en quoi ce but ne peut pas être atteint par le simple recyclage d'ontologies ou de terminologies existantes, ni même de thésaurus utilisés dans le cadre du traitement documentaire classique de l'information. Les résultats obtenus seront discutés au fil des développements.

2. Critères pour la construction d'un RTRI

Nous proposons ici un bref aperçu d'une réflexion en cours concernant la définition d'un cadre explicite pour parvenir au but qui nous occupe. Voici donc un certain nombre d'hypothèses et de postulats sur lesquels s'appuie notre démarche.

Chaque intranaute est détenteur à des degrés variables d'un référentiel d'entreprise. Pour le définir, nous partons de la définition que Josette Rey-Debove (1998 : 289) donne du référentiel :

> « *Ensemble de tout ce dont un locuteur peut parler dans une langue donnée (objets réels ou imaginaires, concrets ou abstraits, appelés référents).* »

Nous en restreignons ensuite le champ en ajoutant que ces objets doivent entretenir un rapport direct ou indirect avec l'accomplissement des tâches pour lesquelles le locuteur est rémunéré. Ce référentiel possède différentes strates : le métier, les domaines qui lui sont liés, jusqu'aux orientations stratégiques de l'entreprise. Il est composé également de différents registres, allant des échanges informels aux textes de communication institutionnelle (journaux internes, rapports annuels, communiqués de presse, etc.).

Les informations diffusées sur un intranet et les recherches que l'on peut y faire relèvent du « parler d'entreprise », tel que l'a défini Dardo de Vecchi[5]. Cet ensemble de termes propres à une entreprise est réparti sur un *continuum*, allant des jargons à la langue commune en passant par différentes terminologies.

Les usagers d'un intranet sont confrontés quotidiennement à des recherches d'informations. Celles-ci varient en fonction du profil de l'intranaute, de son activité, du besoin d'information qu'est censé satisfaire la recherche, de la nature des informations recherchées, de l'image que

[4] Cela nous permet de nous inscrire dans une plus vaste famille d'outils tout en nous démarquant d'une acception trop recentrée sur les langages documentaires, telle celle proposée par Philippe Lefèvre: « *Aujourd'hui, la notion de référentiel terminologique se substitue progressivement à celle de langage documentaire. Sous ce concept, se placent les thésaurus, les réseaux sémantiques, et les terminologies structurées (par exemple incluant des relations de synonymie.* » (Lefèvre, 2000 : 136). Le RTRI n'est pas réductible à un langage documentaire qui « (…) *est un langage artificiel, un métalangage, constitué de notions et de relations entre ces notions. Sa finalité est de formaliser à la fois les notions contenues dans les documents et l'expression des demandes d'informations. C'est un système de représentation synthétique du contenu des textes*. » (Cacaly *et al.*, 1997).

[5] « *Ensemble des termes propres à une entreprise et qui la distingue des autres entreprises.* » (Vecchi, 2002).

les intranautes se font de l'objet de leur recherche et du dispositif de recherche d'informations qu'ils utilisent, etc. Dans la mesure où des analogies sont identifiables entre les caractéristiques de certains discours[6] et celles d'une recherche type, nous faisons l'hypothèse que les schématisations produites dans les deux cas ont des chances d'être similaires.

Les éléments du référentiel de l'intranaute ainsi que les thèmes présents dans les discours produits et diffusés sur l'intranet sont le résultat de schématisations. Celles-ci sont appréhendées dans les corpus sous la forme d'objets de discours, c'est-à-dire d'objets construits par des moyens et des processus linguistiques. Cela laisse la possibilité de prendre en considération les discours oraux aussi bien qu'écrits. Les premiers demandant pour l'instant des moyens d'exploitation hors de notre portée (interviews et transcriptions), par facilité, nos corpus se composent essentiellement de textes au format électronique diffusés sur un intranet.

Notre démarche s'inscrit dans le cadre d'une linguistique de discours qui est associée aux travaux de Zellig Harris (1951 et 1969) et que Rostislav Kocourek (1991 : 24) a plus récemment résumée au sujet de la méthode terminologique :

> « *On peut formuler une* ***hypothèse*** *qu'il est linguistiquement légitime de choisir un ensemble de textes, délimités d'une manière externe, dans le but de déterminer les ressources linguistiques sous-tendues, de dégager les propriétés, les principes, les tendances de ce sous-ensemble de textes, et, en ce faisant, d'enrichir, de préciser, d'approfondir la connaissance et la compréhension de la langue entière.* »

Un choix de textes doit ensuite être organisé en corpus, au sens où le définit Benoît Habert (2000), de manière à pouvoir être exploité avec les garanties et la rigueur nécessaires.

Il convient de préciser également que nous adoptons alternativement les points de vue sémasiologique et onomasiologique pour collecter les unités langagières et conduire l'exploration de leur sens sur les axes syntagmatique et paradigmatique. Par ailleurs, nous recourrons au formalisme utilisé en Sciences de l'information et en Terminologie pour caractériser certaines relations sémantiques[7]. Pour la construction des RTRI, nous nous appuyons sur des logiciels de lexicométrie[8]. Dans la mesure où les traitements qu'ils offrent reposent exclusivement sur la manipulation des formes graphiques, ils se situent en amont des exigences du traitement automatique du langage naturel et permettent une souplesse et une rapidité de mise en place adaptées au but que nous poursuivons. Ils constituent ainsi une approche complémentaire à celle du TALN[9] pour l'extraction d'informations, la constitution de lexiques, d'ontologies ou de terminologies, dans des contextes variés.

---

[6] Schéma de communication proposé par Grize (1996) et Adam (1999) : *« Une schématisation comporte au moins six types d'images de base qui sont proposées par le discours et sont autant de sortes de versions du monde :*

– des images de la situation d'interaction sociodiscursive en cours ;

– des images de l'objet de discours (que l'on appellera aussi bien thème que référent) ;

– des images de A (schématiseur) ;

– des images de B (co-schématiseur).

À ces quatre cas répertoriés par Grize, il faut certainement ajouter encore des images de la langue de l'autre ou de celle que l'autre attend que l'on produise. Cette question fondamentale qui traverse les études sociolinguistiques et la réflexion de Pierre Bourdieu sur le 'capital linguistique' des sujets s'étend également jusqu'aux images du support et/ou du canal de transmission de la schématisation. » (Adam, 1999 : 107). Nous avons retiré de la citation les abréviations utilisées dans un graphique que nous ne reproduisons pas ici.

[7] *Travail terminologique. Principes et méthodes*. NF ISO 704-2001 et *Documentation – Principes directeurs pour l'établissement et le développement de thésaurus monolingues,* ISO 2788-1986 (F).

[8] Il s'agit ici de Lexico version 3 (Salem *et al*.) et d'Hyperbase version 5.1 (Brunet).

[9] Parmi les nombreux travaux ayant adopté cette approche on peut citer ceux de D. Bourigault et C. Jacque-

3. Une réponse aux difficultés de la recherche d'informations sur un intranet

3.1. *Des procédures lexicométriques pour améliorer la recherche d'informations*

Une difficulté majeure de la recherche d'informations sur un intranet provient de l'hétérogénéité des informations mises en ligne. Elle est en quelque sorte redoublée par l'assemblage de systèmes de recherche différents possédant parfois des comportements peu compatibles entre eux. Pour illustrer ces deux points, nous présenterons quatre contextes[10] d'utilisation de l'objet « livret jeune ».

L'exploitation des trois premiers corpus (secteurs juridique, documentaire et comptable) a été réalisée manuellement. En revanche, pour la partie « communication institutionnelle », une série de procédures semi-automatisées a été appliquée à l'aide du logiciel Lexico 3.

3.1.1. *Repérage des formes et des variantes*

Le syntagme « livret jeune » correspond à un terme dans le domaine bancaire. Pour l'exploitation d'un corpus, nous nous appuyons sur la segmentation initiale en formes graphiques réalisée par le logiciel. Aucune normalisation n'est appliquée sur les variantes graphiques ou morphologiques des unités, afin d'assurer un traitement homogène pour les différents corpus utilisés, quelles que soient les contraintes (de temps, de taille, de format de fichier, etc.), pesant sur leur mise en œuvre. Il est donc nécessaire d'envisager toutes les graphies pouvant être utilisées dans les textes pour écrire « livret jeune ». La première procédure consiste à écrire les expressions régulières adéquates servant à interroger le corpus, puis à regrouper les formes obtenues au sein d'un type généralisé (Tgen pour la suite) (Lamalle et Salem, 2002).

Expressions régulières	Liste des formes obtenues regroupés en Tgen
[Ll]ivret|LIVRET	**Tgen « livret » :** livrets (13 occ.) ; livret (8 occ.) ; Livrets (4 occ.) ; Livret (3 occ.)
[Jj]eune|JEUNE	**Tgen « jeune » :** jeunes (57 occ.) ; Jeunes (7 occ.) ; JEUNES (4 occ.) ; jeune (4 occ.) ; Jeune (1 occ.)

Tableau 1. Expressions régulières et Tgen

Chaque Tgen est ensuite associé à une couleur distinctive et projeté sur la carte des sections de Lexico. L'unité utilisée est le paragraphe original. En effet, celui-ci semble donner les meilleurs résultats pour des explorations réalisées dans approche distributionnelle de la sémantique. La carte des sections permet ensuite d'accéder visuellement à tous les contextes (bicolores), où des formes appartenant aux deux Tgen sont cooccurrentes :

---

min (Bourigault, 2000).

[10] Le contexte juridique est représenté par un petit corpus de 2000 occurrences composé des articles L 221-24 à L 221-26 du Code monétaire et financier et du Décret n° 96-367 du 2 mai 1996 relatif à la mise en place du « livret jeune ». Le secteur documentaire est illustré à l'aide du thésaurus RESAGRI (édition 1997), qui se compose de descripteurs utilisés pour réaliser une indexation manuelle des informations. Un échantillon de référentiels internes a été prélevé pour le domaine comptable. Dans tous les cas, il s'agit de corpus ad hoc n'ayant bénéficié que d'un traitement léger, voire nul pour l'extrait de thésaurus. Enfin, pour le secteur de la communication institutionnelle, un corpus de 163 000 occurrences a été utilisé. Celui-ci, contrairement aux trois autres, a fait l'objet d'une construction méticuleuse. Son texte a été contrôlé, balisé, pour identifier parties, rubriques et paragraphes ; les zones de texte hétérogènes comportant des tableaux, graphiques ou autres organigrammes, ont fait l'objet d'une procédure de transposition appliquée systématiquement. Il se compose d'une série textuelle chronologique regroupant huit rapports d'activité d'une banque française. Ces documents présentent une synthèse annuelle de l'activité de la banque à destination du marché.

« *Conformément à la tendance générale, la baisse des taux de marché, la création du* ***livret*** ***Jeunes****, les baisses sélectives des taux des produits réglementés et les mesures fiscales ont induit des changements importants dans la structure de la collecte.* »

Figure 1. Carte des paragraphes originaux et contenu d'un paragraphe

Il devient alors possible d'identifier rapidement toutes les variantes graphiques de « livret jeune » dans ce corpus[11] : « **Livrets Jeunes** » (2 occ.), « **livrets jeunes** » (1 occ.), « **livret Jeunes** » (2 occ.), « **livrets Jeunes** » (2 occ.), « **Livret Jeune** Mozaïc » (1 occ.), « **Livret Jeunes** » (1 occ.).

Les neuf occurrences recensées permettent d'identifier six variantes dans l'usage de la majuscule, deux variantes dans l'usage du pluriel pour « jeune » et une expansion (« Livret Jeune Mozaïc »). Ces aspects graphiques ne doivent pas être négligés dans un environnement où plusieurs moteurs de recherche peuvent être sollicités pour répondre à une même requête. En effet, certains moteurs rudimentaires sont sensibles à la casse et à l'accentuation, ce qui peut augmenter le silence dans la réponse fournie par le système.

La procédure mise en place pour la constitution d'un référentiel prévoit de compléter ce premier examen par l'observation des segments répétés présents dans le corpus. Pour le corpus de communication, le repérage ayant pu être conduit de façon exhaustive dans l'étape précédente, le calcul des segments ne s'impose pas (la simple lecture des contextes permet de dénombrer facilement deux occurrences à chaque fois pour « Livrets Jeunes », « livrets Jeunes » et « livrets Jeunes »). Cependant, il faut mentionner que ce calcul est un très bon indicateur du souci de cohérence et d'homogénéité qui a présidé à la rédaction d'un texte. En effet, dans le corpus juridique, composé de seulement deux mille occurrences, le calcul des segments répétés permet d'attester instantanément l'usage exclusif de la graphie « livret jeune » (vingt-cinq occurrences). Cette étude des données segmentales peut être ensuite affinée à l'aide des concordances afin de prolonger l'exploration de l'axe syntagmatique.

Le recensement des variantes constitue un gain intéressant dans le domaine de la recherche d'informations, puisque nous possédons ainsi la possibilité de créer des équivalences qui, injectées dans la recherche par une expansion de requête, vont assurer un meilleur rappel. Après ce premier traitement, la constitution d'un RTRI nécessite que soient ensuite explorées les relations sémantiques que le syntagme « livret jeune » peut entretenir avec d'autres unités. Pour ce faire, nous recherchons tout vocable utile pour éclairer ce qui relie l'objet « livret jeune » à l'activité d'une banque.

3.1.2. *Analyse des cooccurrences et appartenance thématique*

Le calcul des cooccurrents d'une forme est classiquement utilisé pour explorer le thème auquel une unité se rattache dans les discours. La méthode utilisée par le logiciel repose sur le calcul des spécificités (Lafon, 1984 ; Lebart et Salem, 1994 : 171 et suiv.). Celui-ci compare deux ensembles de vocabulaire correspondant, d'une part, aux contextes où l'unité polylexicale « livret jeune » apparaît, et d'autre part, aux contextes où elle est absente. Pour les formes partageant les mêmes contextes que l'une au moins des variantes graphiques de

[11] La présence de majuscules ne peut pas être expliquée par la position des expressions en début de phrase, aucune des occurrences recensées n'occupant cette place.

« livret jeune », ce calcul propose un diagnostic indiquant si elles s'y trouvent en sous-emploi ou en sur-emploi. Le coefficient de la colonne « Coeff. » donne le degré de spécificité de la forme. Seules les premières spécificités positives ont été retenues ici, car elles correspondent aux cooccurrents les plus significatifs. Enfin, le calcul a été lancé avec les paramètres suivants : fréquence des formes prises en compte supérieure ou égale à 2 et seuil de probabilité égal à 5 %[12].

Forme	Frq. Tot.	Fréquence	Coeff.
Mozaïc	31	11	20
Jeunes	7	7	18
carte	58	6	8
livrets	13	4	8
prêt	11	3	6
livret	8	3	6
jeunes	57	5	6

Forme	Frq. Tot.	Fréquence	Coeff.
JEUNES	4	2	5
Livret	3	2	5
étudiant	3	2	5
Livrets	4	2	5
vis[13]	4	2	5
MOZAÏC	2	2	5

Tableau 2. Les treize premiers cooccurrents des différentes graphies de « livret jeune »

Ce résultat permet d'effectuer une première série de constats. Tout d'abord nous remarquons la présence de la forme « Mozaïc » avec deux graphies différentes. Nous avons vu plus haut que cette forme peut constituer une expansion de « livret jeune ». Il convient donc d'examiner plus en détail la nature de la relation qui la lie à la dénomination « livret jeune ». Nous retrouvons par ailleurs un certain nombre de formes appartenant aux deux Tgen. Mais les indications chiffrées nous interpellent sur le fait que pour « jeune » comme pour « livret », le nombre d'occurrences est bien supérieur à celui utilisé pour toutes les variantes de « livret jeune » dans le corpus. Il devient alors nécessaire de savoir à quoi renvoient ces unités en dehors de leur emploi dans le syntagme lexicalisé qui nous a servi de pôle. Nous pouvons noter cependant que cet écart vaut surtout pour les formes sans majuscule, car celles qui en possèdent une sont presque toutes absorbées par les variantes de « livret jeune » (c'est le cas pour toutes les occurrences de la graphie « Jeunes », de deux sur trois au total pour « Livret » et de deux sur quatre pour « Livrets »). Cet usage sélectif de la majuscule n'est certainement pas le fruit du hasard, et il nous faudra commenter ce point plus loin. Enfin, un quatrième ensemble constitué des formes « carte, prêt, étudiant et vis [à vis] », semble suggérer que le « livret jeune » fait partie d'un ensemble plus vaste de produits bancaires. À ce stade de l'exploration, il semble difficile de pousser plus loin le commentaire, à moins d'explorer chaque contexte ou d'affiner à nouveau notre analyse au moyen du calcul des cooccurrents.

3.1.3. *Calcul des cooccurrents de cooccurrents et affinage des informations thématiques*

Cette méthode a été proposée par Martinez (2000) pour la recherche des synonymes d'une forme pôle. Nous l'étendons ici à une forme d'exploration plus large, où elle nous semble produire également des résultats significatifs dans la mise à jour des différentes facettes d'un

[12] La fréquence supérieure ou égale à 2 a été retenue compte tenu des fréquences peu élevées sur lesquelles portent les calculs en général pour l'exploitation de ce corpus. Quant au seuil de 5 %, il constitue un bon compromis entre filtrage et sélection des formes spécifiques.

[13] Il s'agit de « vis à vis », dans deux phrases :
P1 : « Depuis plusieurs années , le Crédit agricole mène une politique dynamique **vis**-à-**vis** des jeunes . »
P2 : « Une attention particulière **vis**-à-**vis** des jeunes pour préparer l'avenir : »

objet de discours. Le principe en est simple : après avoir établi la liste des cooccurrents d'une forme ou d'un Tgen pôle, un nouveau Tgen est constitué dans lequel la forme pôle est absente (ainsi que ses variantes), mais où sont regroupés ses cooccurrents les plus spécifiques. Le calcul des spécificités est alors réitéré dans le but d'obtenir les cooccurrents des cooccurrents de la forme pôle. Cette procédure revient donc à rechercher les unités qui partagent un environnement similaire à celui de la forme pôle mais où celle-ci n'apparaît pas forcément. L'observation des objets nouveaux que le calcul fait apparaître est susceptible d'éclairer la nature des relations tissées dans le corpus entre la forme pôle et des unités comme « Mozaïc », « jeunes » ou « carte ».

Constitution d'un Tgen « cooc. livret jeune » avec les 5 premiers cooccurrents de « livret jeune » en dehors des formes contenues dans les Tgen « livret » et « jeune », puis calcul des formes cooccurrentes :

Forme	Frq. Tot.
Mozaïc	31
carte	58
prêt	11
étudiant	3
MOZAÏC	2

Tgen « cooc. livret jeune »

Principaux cooccurrents du Tgen « cooc. livret jeune »

Forme	Frq. Tot.	Frq. Part.	Coeff.
carte	58	58	50
Mozaïc	31	31	50
Open	20	20	33
prêt	11	11	18
ans	73	21	17
000	174	28	15
cartes	39	14	13
détenteurs	11	9	13
jeunes	57	16	13
paiement	61	13	9
Service	30	10	9
consommation	138	17	8

Forme	Frq. Tot.	Frq. Part.	Coeff.
crédit	314	25	7
Compte	37	9	7
Sofinco	133	16	7
25	158	17	7
Eurocard	4	4	7
Maestro	7	5	7
Jeunes	7	5	7
découvert	6	4	6
avec	790	42	6
18	96	12	6
offre	251	20	6
retrait	10	5	6

Tableaux 3 et 4. Tgen « cooc. livret jeune » et ses cooccurrents

On n'analysera pas ce tableau dans le détail dans le cadre de cet article. Cependant, il est possible de préciser un certain nombre de points laissés en suspens dans la section précédente. Ainsi, l'unité « Mozaïc » apparaît en fait comme la dénomination d'une gamme de produits bancaires destinés aux jeunes, ce que permet d'attester un contexte de ce type :

> « *L'offre pour les jeunes, commercialisée sous la marque* ***Mozaïc****, s'est enrichie en 1998 d' un site Internet dédié. En outre, le Compte-Service* ***Mozaïc****, destiné aux 18-25 ans, a*

été généralisé. Celui-ci est composé de produits répondant aux besoins courants (sécurisation des moyens de paiement, autorisation de découvert) et de différents services complémentaires : ***prêt étudiant****, Livret Jeunes ou assurance habitation.* »

La présence des trois zéros, qui correspondent à la notation conventionnelle du millier, comme dans « 300 000 » par exemple, est significative. Elle correspond au chiffrage d'un certain nombres d'indicateurs de l'activité bancaire, comme le nombre de contrats de service souscrits, de détenteurs de cartes bancaires (cartes de retrait, de paiement ou de crédit, baptisées Eurocard, Mozaïc ou Maestro), de dossiers de prêt enregistrés, d'ouvertures de comptes ou de livrets. Ces éléments permettent de rattacher sans ambiguïté le « livret jeune » aux produits bancaires classiques proposés aux jeunes. D'autres chiffres sont également révélateurs : il s'agit de « 25 » et de « 18 » qui, combinés avec la forme « ans », forment l'unité « 18 – 25 ans » qui délimite le segment de clientèle auquel les produits évoqués plus haut sont proposés.

Il reste que seulement un tiers des occurrences de « jeunes » (16 occurrences, ou 17 si l'on tient compte de l'occurrence mentionnée plus haut à propos de la graphie « livrets jeunes », sur les 57 que compte la forme dans l'ensemble du corpus), a été attiré par les formes rassemblées dans le Tgen « cooc. livret jeune ».On peut alors faire l'hypothèse que cette forme appartient à plusieurs thèmes qu'il s'agit d'identifier.

Dans l'exemple précédent, la réitération du calcul des spécificités a été réalisée à partir d'un ensemble de formes regroupées au sein d'un Tgen, afin de repérer les environnements semblables à ceux dans lesquels apparaît l'unité polylexicale « livret jeune ». Maintenant, nous souhaitons, à l'inverse, identifier des environnements thématiques différents dans lesquels la même forme « jeune » est présente. Le Tgen « jeune » étant constitué avec toutes les variantes graphiques et morphologiques de « jeune » présentes dans le corpus (cf. section 3.1.1), nous calculons ses cooccurrents.

Forme	**Frq. Tot.**	**Frq. Part.**	**Coeff.**
jeunes	57	57	50
Mozaïc	31	24	33
ans	73	25	22
installation	17	13	18
carte	58	16	13
Jeunes	7	7	12
alternance	8	7	11
apprentissage	8	7	11
transmission	12	8	11

Tableau 5. Principaux cooccurrents du Tgen « jeune »

Dans une deuxième étape, nous procédons au calcul et à l'observation des cooccurrents des principaux cooccurrents du Tgen « jeune », excepté les formes « jeunes » et « Jeunes ».

La réitération du calcul des spécificités fait apparaître sept listes de cooccurrents qui peuvent être regroupées en quatre ensembles assez cohérents (nommés A, B, C et D), compte tenu des formes qu'ils partagent. Ces ensembles éclairent chacun une facette différente du Tgen « jeune ». Dans l'ensemble « A », le segment de clientèle des jeunes se manifeste nettement avec les produits qui lui sont proposés et les appellations « 18-25 ans » ou « moins de 25 ans ». La liste « B » livre les ingrédients d'une thématique ayant pour centre la carte bancaire (carte Mozaïc de paiement, carte de crédit à la consommation Open, carte Maestro, etc.), et dans laquelle les jeunes semblent être d'abord assimilés à des détenteurs de cartes. Enfin, les

deux ensembles « C » et « D » suggèrent l'existence dans le corpus de deux sous-populations spécifiques de jeunes. Dans l'ensemble « C » il s'agit de celle des jeunes agriculteurs (leur installation, la transmission des exploitations agricoles, etc.). Enfin, l'ensemble « D » semble résumer la thématique du parcours de certains jeunes débutant dans la vie active (apprentissage, formation continue, etc.).

Forme	**Forme**	**Forme**	**Forme**	**Forme**	**Forme**	**Forme**
Mozaïc	ans	carte	installation	transmission	alternance	apprentissage
carte	Mozaïc	Open	transmission	installation	formation	apprentis
ans	jeunes	Mozaïc	jeunes	jeunes	apprentis	alternance
jeunes	carte	détenteurs	exploitations	agriculteurs	apprentissage	emploi
Jeunes	moins	cartes	accompagnement	exploitations	jeunes	formation
18	25	consommation	agriculteurs	Clés	contrats	jeunes
25	depuis	paiement	Clés	notaires	faveur	contrats
Service	catégorie	Sofinco	distributeurs	retraite	emploi	faveur
détenteurs	Comptes	ans	Mozaïc	Sodica	associant	associant
Compte	18	Maestro	Charte	Prediagri	IFCAM	temps
A		B	C		D	

Tableau 6. Les dix premiers cooccurrents des principaux cooccurrents du Tgen « jeune » regroupés en quatre ensembles (A, B, C, et D)

Ces regroupements nécessiteraient naturellement un examen plus poussé pour alimenter un RTRI. On constate, néanmoins, qu'ils contribuent rapidement et efficacement à élucider l'énigme concernant les différents contextes d'utilisation du Tgen « jeune ».

3.2. *RTRI et prise en compte de stratégies de dénomination concurrentes*

Le tableau ci-dessous présente, de façon schématique, différents contextes d'apparition de l'objet « livret jeune », avec leurs dénominations et leurs catégorisations spécifiques.

JURIDIQUE Titre II Les produits d'épargne (…) Chapitre 1 Produits d'épargne générale à régime fiscal spécifique (…) Section 3 Le livret jeune Dénominations : livret jeune / livrets jeunes	DOCUMENTAIRE ressources bancaires (…) épargne (…) livret d'épargne livret d'épargne jeune Dénomination : livret d'épargne jeune
COMPTABLE Collecte Collecte épargne (…) Épargne livret (…) Livrets jeunes Dénominations : Livrets jeunes / Livret Jeune Mozaïc / LMZ	COMMUNICATION INSTITUTIONNELLE Liste des principales formes cooccurrentes caractérisant les paragraphes où « livret jeune » apparaît : Mozaïc ; Jeunes ; carte ; livrets ; prêt ; livret ; jeunes ; JEUNES ; Livret ; étudiant ; Livrets ; vis ; MOZAÏC . Dénominations : livrets Jeunes / livrets jeunes / livret Jeunes / Livrets Jeunes / Livret Jeune Mozaïc / Livret Jeunes

Tableau 7. Sites intranets et usages langagiers

Lorsque nous comparons les dénominations recensées dans les quatre secteurs mentionnés plus haut, deux faits ressortent particulièrement. Tout d'abord, la dénomination officielle « livret jeune » observée dans le Code monétaire et dans le Décret d'application ne se

retrouve pas dans les autres secteurs, ou tout au moins avec sa graphie exacte. Notons que cette répartition reflète la séparation entre extérieur et intérieur par rapport l'entreprise, alors que l'on peut dire que toutes les sources recensées ont un rapport assez étroit avec l'activité bancaire. D'autre part, il faut bien constater qu'entre secteurs d'une même entreprise la situation n'est guère meilleure, puisque l'échantillon représentant l'un d'entre eux ne possède aucune forme commune avec les autres. On va voir que cela peut s'expliquer par la nature des utilisations qui sont faites de l'objet « livret jeune ». La prise en compte de ce phénomène constitue un enjeu très important pour la mise en place d'un RTRI.

La dénomination « livret d'épargne jeune », présente dans le secteur documentaire, est ignorée, entre autres, par les textes juridiques qui représentent pourtant une source on ne peut plus légitime. De ce point de vue, il s'agit d'une forme artificielle créée pour la construction d'un thésaurus et destinée à l'indexation documentaire. Le thésaurus est, en effet, une construction hiérarchisée de termes qui doivent représenter non seulement le domaine, mais aussi son organisation. Celle-ci s'incarne en particulier dans le mécanisme « terme de tête – expansion ». Le livret jeune étant un livret d'épargne, il est admis qu'un terme artificiel puisse être créé dans le seul but de rendre explicite cette information.

C'est un peu le même mécanisme qui caractérise le faux sigle « LMZ », pour « livret Mozaïc ». Cette dernière dénomination n'est pas non plus répertoriée. En revanche, il existe d'autres sigles à trois lettres concernant les livrets d'épargne : LEE, pour le livret d'épargne entreprise et LEP, pour le livret d'épargne populaire. Un sigle conforme aux usages « locaux » a donc été créé de toutes pièces : il comporte trois lettres, il est facile à prononcer et évocateur, qualités absentes de « LJ ».

Les flottements les plus remarquables au sujet du « livret jeune » se rencontrent cependant dans le corpus de communication institutionnelle. Ils revêtent deux aspects : l'incertitude sur l'attribution de la majuscule à « livret » et à « jeunes » et l'utilisation quasi systématique du pluriel pour « jeunes », quand bien même « livret » se trouve être au singulier. Le livret jeune s'adresse à une clientèle de jeunes (l'analyse des spécificités a permis d'identifier deux facettes importantes du produit d'épargne : son appartenance à une gamme de produits destinés aux jeunes – Mozaïc –, et les liens étroits qu'il entretient avec un segment de ce marché). Aussi, la dénomination « livret jeune » est-elle nécessairement amenée à côtoyer dans les textes celle de « jeunes ». Compte tenu de ces éléments, nous interprétons la majuscule comme la marque du souci de délimiter la dénomination du produit par rapport à son cotexte. Reste à rendre compte du pluriel qui est appliqué aussi quand « livret » est au singulier. L'explication nous semble liée à ce que nous venons de dire. En effet, pour le législateur, le livret jeune se comprend comme un livret unique dont peut être titulaire « une personne physique âgée de douze à vingt-cinq ans et résidant en France à titre habituel », ce qui correspond à une définition possible. En revanche, dans le contexte bancaire, « jeune » au singulier est presque ignoré, car ce qui importe, c'est le segment de clientèle, le marché des jeunes. Dans ce contexte, il se comprend donc comme le livret d'épargne destiné aux jeunes.

Si on tient compte des contextes pris à titre d'exemple sur un intranet, un réseau sémantique de ce type peut être élaboré (Fig. 2).

4. Conclusion

Cette étude a permis de mettre l'accent sur la spécificité du référentiel terminologique dédié à la recherche d'informations. Celui-ci se compose d'unités et de relations sémantiques en usage dans un sociolecte et susceptibles d'être utilisées pour des recherches d'informations menées dans un cadre précis. Cette dernière caractéristique impose les contraintes suivantes :

une mise en œuvre rapide, des mises à jour régulières, le traitement de corpus de taille et de qualité variables, le recours à des genres de textes hétérogènes, etc. C'est pourquoi, il nous semble que les tâches lexicographiques ou terminographiques impliquées par la constitution d'un RTRI requièrent des logiciels d'une grande portabilité, et orientés à titre principal vers l'exploration des données. Dans ce domaine, la navigation hypertextuelle (Hyperbase), ou les possibilités de *drag and drop*, de cartographie textuelle ou de constitution de types généralisés (Lexico), représentent un avantage décisif.

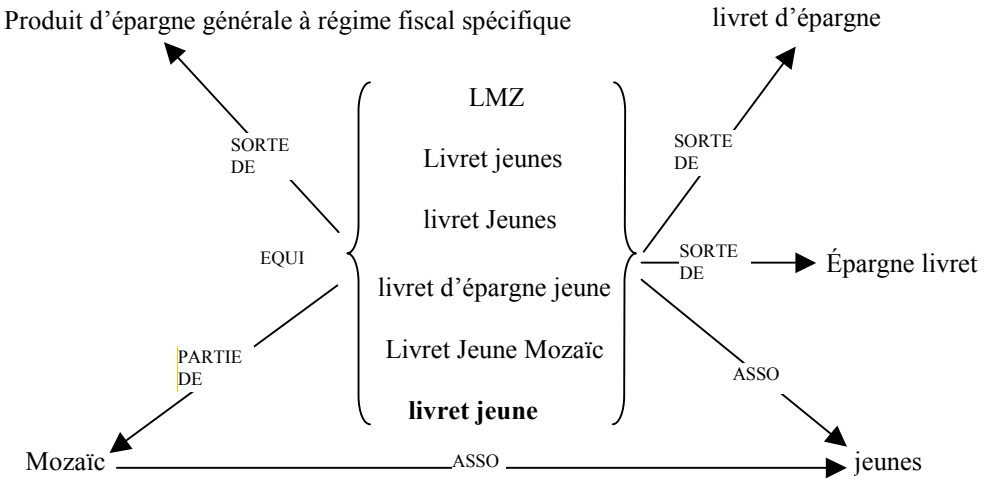


Figure 2. Réseau sémantique de « livret jeune » dans un RTRI[14]

Ces outils permettent en effet d'obtenir des indications précises sur la graphie, les homonymes ou les variantes morphologiques d'une unité grâce aux index disponibles, aux fonctionnalités de concordances, inventaires distributionnels et cooccurrences. La question délicate de l'identification des syntagmes lexicalisés trouve une réponse de premier niveau avec les segments répétés. L'exploration de la sémantique lexicale est grandement facilitée par la fourniture d'indications factuelles sur les régularités distributionnelles observables à l'intérieur de fenêtres dont la taille peut varier en fonction des besoins.

Dans cette étude nous nous sommes limité à une unité terminologique typique afin d'illustrer l'application de ces programmes à un contexte de parler d'entreprise, où langue commune, langues spécialisées et jargons sont étroitement mêlés. À l'option représentée par une homogénéisation de corpus *a posteriori*, nous avons préféré la constitution de corpus sous la forme de séries chronologiques, caractérisées par des situations d'énonciation homogènes. Cependant, un certain nombre de difficultés demeurent, inhérentes aux conditions d'exploitation de ces matériaux. Les calculs statistiques utilisés pour les fréquences relativement élevées ne sont pas appliqués aux fréquences trop faibles. Réciproquement, l'exploration en contexte d'un nombre restreint d'occurrences est utilisée de façon moins systématique pour les formes fréquentes. Dans la mesure où des traitements différents semblent inévitables, la nature de l'unité incarnée par la forme doit toujours venir pondérer l'interprétation. Ce type de contrôle est d'autant plus nécessaire que le comportement d'une même unité peut être très variable

[14] Les relations entre unités sont encadrées et correspondent : « ASSO » à une relation associative, « EQUI » à une relation d'équivalence, « SORTE DE » à la relation hiérarchique du même nom, « PARTIE DE » à la relation hiérarchique du même nom. Ces relations classiques font l'objet d'une définition dans les normes précitées. La flèche indique ici le sens de la relation. Mais la relation entre deux unités est bi-directionnelle, c'est-à-dire que « PARTIE DE » dans un sens, se lit « A POUR PARTIE » dans l'autre sens.

d'un corpus à l'autre. Par exemple, un indice de lexicalisation des syntagmes nominaux, adopté pour un corpus, doit souvent être révisé pour un autre, si l'on veut éviter qu'il ne devienne trop ou pas assez filtrant. L'exploration du « défricheur » doit donc assez rapidement faire place à celle du « tacticien », ce qui peut demander un apprentissage plus long que celui qui est requis pour la manipulation des logiciels eux-mêmes.

Références

Adam J.-M. (1999). *Linguistique textuelle - Des genres de discours aux textes*. Nathan/HER, coll. fac.

Amar E. (1997). *Internet-intranet – Les Concepts de base*. AFNOR, vol. (1-2).

Bommier-Pincemin B. (1999). *Diffusion ciblée automatique d'information : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thèse, Université de Paris IV.

Bourigault D. et Jacquemin C. (2000). Construction de ressources terminologiques. In Pierrel J.-M. (Ed.), *Ingénierie des langues*. Hermès : 215-230.

Brunet Ét. (2001). *Hyperbase Logiciel documentaire et statistique pour la création et l'exploitation de bases hypertextuelles, Manuel de référence, version 5.2*, CNRS, Université de Nice, 1999, 2001 (pour la mise à jour).

Cacaly S. (sous la direction de) (1997). *Dictionnaire encyclopédique de l'information et de la documentation*. Nathan.

Grize J.-B. (1996). *Logique naturelle et communications*. PUF.

Habert B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment ? *Linguistiques sur corpus. Études et réflexions, Cahiers de l'Université de Perpignan*, vol. (31), PUP : 11-58.

Harris Z.S. (1951). *Methods In Structural Linguistics*. The University of Chicago Press.

Harris Z.S. (1969). (trad. de l'américain par F. Dubois-Charlier). Analyse du discours. *Langages*, vol. (13). Didier – Larousse : 8-45 [1ère édition : *Language,* vol. (28), 1952 : 1-30].

Kocourek R. (1991). *La Langue française de la technique et de la science – Vers une linguistique de la langue savante*. Wiesbaden, Brandstetter Verlag.

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Slatkine-Champion.

Lamalle C. et Salem A. (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. In *Actes des JADT 2002*.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Dunod.

Lefèvre P. (2000). *La Recherche d'informations - Du texte intégral au thésaurus*. Hermès.

Leselbaum J. et Labbé D. (2002). Lexicographie assistée par ordinateur. Signification de « Banque » dans le vocabulaire économique. In *Actes des JADT 2002* : 447-458.

Martinez W. (2000). Mise en évidence de rapports synonymiques par la méthode des cooccurrences. In *Actes des JADT 2000*.

Mortureux M.-F. (1997). *La Lexicologie entre langue et discours*. SEDES.

Otman G. (1996). *Les Représentations sémantiques en terminologie*. Masson.

Thesaurus RESAGRI 1997, 4 vol. alphabétique, permuté, thématique et géographique. RESAGRI.

Rey A. (1992). *La Terminologie noms et notions*. PUF, coll. Que sais-je ?

Rey-Debove J. (1998). *La Linguistique du signe. Une approche sémiotique du langage*. Armand Colin.

Salem A. (1987). *Pratique des segments répétés – Essai de statistique textuelle*. Klincksieck.

Salem A. (1993). *Méthodes de la statistique textuelle*, Thèse pour le doctorat d'État, Université de Paris III Sorbonne Nouvelle.

Vecchi de D. (2002). *Vous avez dit jargon…*Eyrolles.

A Simple LNRE Model for Random Character Sequences

Stefan Evert

IMS – University of Stuttgart – Azenbergstr. 12 – 70174 Stuttgart – Germany
evert@ims.uni-stuttgart.de

Abstract

This paper describes a population model for word frequency distributions based on the Zipf-Mandelbrot law, corresponding to the word frequency distribution induced by a random character sequence. The model, which has convenient analytical and numerical properties, is shown to be adequate for the description of language data extracted by automatic means from large text corpora. It can thus be used to study the problems faced by the statistical analysis of such data in the field of natural-language processing.

Keywords: lexical statistics, LNRE models, Zipf-Mandelbrot law, random text, cooccurrence statistics.

1. Introduction to lexical statistics and LNRE models

Most work in the area of lexical statistics is based on random sampling with replacement.[1] This model assumes a population of types $w_1, \cdots, w_S$ with occurrence probabilities $\pi_1, \cdots, \pi_S$. S is called the population size and may be infinite ($S = \infty$) in the case of a countably infinite population. The probabilities π_i are the parameters of this model and must satisfiy

$$\pi_1 + \cdots + \pi_S = 1. \tag{1}$$

It is convenient to assume that they are arranged in descending order, i.e. $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_S$. The random selection of a token from this population is described by a random variable $X : \Omega \to \{1, \ldots, S\}$.[2] A value of $X = k$ implies that the selected token is of type w_k, and the distribution of X is given by $P\big(X = k\big) = \pi_k$ for $k \in \{1, \ldots, S\}$. A random sample of size N corresponds to the N-fold independent repetition of this experiment, i.e. to independent random variables $\mathbf{X} = (X_1, \ldots, X_N)$ with distributions identical to that of X.

In lexical statistics, a text sample of N tokens (which may be anything ranging from orthographic words, over words belonging to a specific morphological category, to word pairs representing cooccurrences) is interpreted as such a random sample $\mathbf{X}$. In this view, two major goals of the statistical analysis are: (i) Draw inferences about the population parameters from the observed data, which are then interpreted in light of the research question. An example is the estimation of the population size S, which may correspond to the number of different word types that a particular word-formation process can generate (*e.g.* Baayen, 2001: Sec. 6.2.) or to the size of an author's vocabulary (*e.g.* McNeil, 1973). (ii) Given the estimated population parameters (or, more generally, assumptions about these parameters), predict the behaviour of various observable quantities. Such quantities correspond to random variables in the random

[1] For all the concepts and results introduced in this section, see Baayen (2001). The notation has been adopted from the same source with minor changes.

[2] When $S = \infty$, $\{1, \ldots, S\}$ stands for the set $\mathbb{N}$ of all natural numbers.

sample model, for which expectations and variances can be computed. A typical example is the prediction of vocabulary growth curves, which measure the increase in the number of observed types when the sample size N is increased (see Baayen, 2001).

Since the random variables $(X_1, \ldots, X_N)$ are jointly independent, the sequential ordering of the tokens in the sample $\mathbf{X}$ provides no information about the population parameters.[3] It is therefore sufficient to consider the type frequencies f_i for $i \in \{1, \ldots, S\}$ (in the mathematical terminology, $\mathbf{f} = (f_1, \ldots, f_S)$ is a sufficient statistic for the random sample $\mathbf{X}$). The frequency f_i is the number of tokens in the sample $\mathbf{X}$ belonging to type w_i, or formally

$$f_i := \sum_{k=1}^{N} I_{[X_k = i]}, \tag{2}$$

where

$$I_{[X_k = i]} := \begin{cases} 1 & X_k = i \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

is the general notation for indicator variables (= random variables with range $\{0, 1\}$). Each f_i is a binomially distributed random variable with success probability π_i, i.e.

$$P\big(f_i = k\big) = \binom{N}{k} (\pi_i)^k (1 - \pi_i)^{N-k} \tag{4}$$

for $k \in \{1, \ldots, N\}$. It has to be kept in mind, though, that the f_i are not mutually independent. The mathematical analysis is considerably simplified when conditioning on a fixed sample size N is avoided, i.e. one assumes that the sample size is itself a Poisson-distributed random variable with mean N. The type frequencies then become *independent* Poisson-distributed random variables with

$$P\big(f_i = k\big) = e^{-N\pi_i} \frac{(N\pi_i)^k}{k!}. \tag{5}$$

This approach (henceforth called independent Poisson sampling) is quite natural when studying the number of different species in biological samples, where the total number of specimens in the sample is obviously subject to random variation and cannot be fixed in advance.[4] Independent Poisson sampling can also be applied in lexical statistics, especially for large N. For instance, the unconstrained sample size with mean $N = 1\,000\,000$ has a standard deviation of $\sigma = \sqrt{N} = 1\,000$; therefore, the observed sample size will almost certainly deviate from N by less than 1%.[5] In the following, I will always assume independent Poisson sampling, as does Baayen (2001). The expectation of f_i is then $E\big[f_i\big] = N\pi_i$, and its variance is $VAR\big[f_i\big] = N\pi_i$.

Since it is usually not known which one of the observed types is the i-th population type w_i, the set of observed frequencies cannot be matched directly against the random variables f_i. It is common practice to arrange the observed frequency values in descending order $f_1^* \geq f_2^* \geq \ldots$, which is called a Zipf ranking. Although π_1 is the highest type probability, and $E\big[f_1\big]$ the highest expectation, w_1 need not be the most frequent observed type corresponding to f_1^*, and

[3] This ordering can be used to test the adequacy of the random sample model, though, e.g. with the dispersion test described in Baayen (2001, Sec. 5.1).

[4] A good deal of the work on word frequency distributions originates in this area, e.g. Good (1953), Holgate (1969), Engen (1974).

[5] The same argument shows that great care has to be taken when Equation (5) is applied to small samples. For $N = 1\,000$, the standard deviation is $\sigma = \sqrt{1\,000} \approx 31.62$ and deviations as far as 10% from the mean N have to be expected.

the same holds for all f_i^*. A better approach is to look at other summary statistics that can be directly observed without an exact knowledge of the population types.

In order to do so, collect all types w_i with the same frequency $f_i = m$ into the frequency class m. The class size V_m, i.e. the number of different types in the frequency class m, can be easily determined from the observed sample. In the random sample model, it is given by the random variable

$$V_m := \sum_{i=1}^{S} I_{[f_i=m]}. \tag{6}$$

The sequence of all class sizes $(V_1, V_2, \dots)$ is called the frequency spectrum. Note that all but finitely many of the V_m equal zero (in particular, the largest non-empty frequency class is $V_{f_1^*}$). Using the same definition, V_0 is the number of unobserved types, which cannot be determined from the sample. The vocabulary size V is the total number of types observed in the sample:

$$V := \sum_{i=1}^{S} I_{[f_i>0]}. \tag{7}$$

The frequency spectrum is related to V and N through the identities $V = \sum_{m=1}^{\infty} V_m$ and $N = \sum_{m=1}^{\infty} m V_m$. The expectations of V and V_m can easily be computed from (5):

$$E\big[V_m\big] = \sum_{i=1}^{S} e^{-N\pi_i} \frac{(N\pi_i)^m}{m!} \quad \text{and} \quad E\big[V\big] = \sum_{i=1}^{S} \left(1 - e^{-N\pi_i}\right), \tag{8}$$

but it is more difficult to obtain variances and the exact distributions (see Baayen, 2001).

As noted before, it is impossible to estimate the large number of probability parameters directly from a sample. It is therefore necessary to formulate a population model with a small number of parameters: once these have been estimated, the hypothesised distribution of the probability parameters π_i can be computed. Following Baayen (2001), I use the term LNRE model for such a population model.[6] While it is in principle possible to formulate an LNRE model directly for the type probability parameters (e.g. Holgate, 1969), it is usually more convenient to use the structural type distribution, which is a step function given by

$$G(\rho) := \big|\{i \in \{1, \dots, S\} \mid \pi_i \geq \rho\}\big|. \tag{9}$$

$E\big[V_m\big]$ and $E\big[V\big]$ can then be expressed in terms of Stieltjes integrals

$$E\big[V_m\big] = \int_0^{\infty} \frac{(N\pi)^m}{m!} e^{-N\pi}\, dG(\pi), \quad E\big[V\big] = \int_0^{\infty} (1 - e^{-N\pi})\, dG(\pi) \tag{10}$$

(Baayen, 2001: 47f). Most LNRE models approximate $G(\rho)$ by a continuous function with type density function $g(\pi)$, i.e.

$$G(\rho) = \int_{\rho}^{\infty} g(\pi)\, d\pi. \tag{11}$$

Note the use of $+\infty$ as an upper integration limit although all type probabilities must fall into the range $0 \leq \pi \leq 1$. This device allows for more elegant mathematical formulations, but care

[6] LNRE stands for Large Number of Rare Events, a term introduced by Khmaladze (1987). It refers to the very large number of types with low occurrence probabilities that are characteristic of word frequency distributions and the associated population models.

has to be taken that $G(1) \ll 1$ (otherwise the LNRE model would predict the existence of types with $\pi > 1$). For an LNRE model based on a type density function $g(\pi)$, the expectations of V_m and V become

$$E\big[V_m\big] = \int_0^\infty \frac{(N\pi)^m}{m!} e^{-N\pi} g(\pi)\, d\pi, \quad E\big[V\big] = \int_0^\infty (1 - e^{-N\pi}) g(\pi)\, d\pi. \tag{12}$$

Equation (1) leads to the normalisation condition

$$\int_0^\infty \pi \cdot g(\pi)\, d\pi = 1, \tag{13}$$

and the population size is given by $S = \int_0^\infty g(\pi)\, d\pi$.

2. Random character sequences and the Zipf-Mandelbrot law

Zipf's law (Zipf, 1949), which states that the frequency of the r-th most frequent type is proportional to $1/r$, was originally formulated for the Zipf ranking of observed frequencies ($f_r^* \approx Cr^{-1}$) and (more or less equivalently) for the observed frequency spectrum ($V_m \approx C/m(m+1)$). In its first form, Zipf's law describes a fascinating property of the higher-frequency words in a language, for which explanations related to Zipf's principle of least effort have been put forward (e.g. Mandelbrot, 1962; Powers, 1998). In its second form, it is a statement about the enormous abundance of lowest-frequency types, which has many consequences for the statistical analysis and for applications in natural-language processing.

It has long been known that the word frequency distributions obtained from random text are strikingly similar to Zipf's law (Miller, 1957; Li, 1992). Formally, random text is understood as a character sequence generated by a Markov process, with word boundaries indicated by a special "space" character. Rouault (1978) shows that, under very general conditions, this segmented character sequence is equivalent to a random sample of words (with replacement, corresponding to the model introduced in Section 1) and that the population probabilities of low-frequency types asymptotically satisfy the Zipf-Mandelbrot law

$$\pi_i = \frac{C}{(i+b)^a} \tag{14}$$

with parameters $a > 1$ and $b > 0$ (Baayen, 2001: 101ff). In Sections 3 and 4, I will formulate LNRE models for random character sequences based on the Zipf-Mandelbrot law. Although Baayen remarks that "for Zipf's harmonic spectrum law and related models, no complete expression for the structural type distribution is available" (Baayen, 2001: 94), this need not discourage us: (14) refers to the population parameters rather than to the observed Zipf ranking. The Zipf-Mandelbrot law for random text is a population model, while the original formulation of Zipf's law and its variants (Baayen, 2001: 94f) have a purely descriptive nature.

These considerations open up an entirely new perspective on Zipf's law: If an LNRE model based on (14) can be shown to agree with the observed data, we must conclude that — as far as statistical analysis is concerned — such language data is not substantially different from random text. As a consequence, the statistical analysis faces all the problems of making sense from random noise, and these problems can be predicted with the LNRE models of Sections 3 and 4.

One of the characteristics of random text is an infinite population size, since there can be words of arbitrary length, leading to an extremely skewed LNRE distribution. It has often been noted

that this does not accord well with real-world data, especially when there are narrow restrictions and the data have been cleaned up manually. Examples are studies of (morphological) productivity (e.g. Baayen and Renouf, 1996) or the word frequency distributions of small literary texts (see Baayen, 2001). However, the situation is different when one considers "raw" data obtained from a large corpus of hundreds of millions of words, which is the input that statistical methods in natural-language processing typically have to deal with. The similarity to random text becomes even more striking for combinations of two or more words (Baayen, 2001: 221). Most techniques for the extraction of collocations from text corpora apply statistical independence tests to such base material (e.g. Evert and Krenn, 2001), and are thus also affected by the consequences of the Zipf-Mandelbrot law. Ha *et al.* (2002) demonstrate such an effect for Mandarin Chinese ideographs: while the number of different graphs is comparatively small and does not exhibit an LNRE distribution, the situation changes when sequences of two or more such graphs are examined. The longer the sequences, the more closely their frequency distribution agrees with the Zipf-Mandelbrot law.

3. The Zipf-Mandelbrot (ZM) LNRE model

In order to derive a useful LNRE model from the Zipf-Mandelbrot law, it is necessary to reformulate (14) in terms of a type density function $g(\pi)$. The structural type distribution corresponding to the Zipf-Mandelbrot law is a step function with $G(\pi_i) = i$ (since there are exactly i types with $\pi \geq \pi_i$, namely $w_1, \ldots, w_i$). Solving (14) for i, we obtain

$$G(\pi) = \frac{C^{1/a}}{\pi^{1/a}} - b \tag{15}$$

for $\pi = \pi_i$, and $G(\pi)$ is constant between these steps. Differentiation of (15) suggests a type density of the form

$$g(\pi) := \begin{cases} C \cdot \pi^{-\alpha-1} & 0 \leq \pi \leq B \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

with two free parameters $0 < \alpha < 1$ and $B > 0$.[7] The normalising constant C can be determined from (13):

$$1 = \int_0^B \pi g(\pi)\, d\pi = \int_0^B C\pi^{-\alpha}\, d\pi = C \cdot \left[\frac{\pi^{1-\alpha}}{1-\alpha}\right]_0^B = C \cdot \frac{B^{1-\alpha}}{1-\alpha} \tag{17}$$

which evaluates to

$$C = \frac{1-\alpha}{B^{1-\alpha}}. \tag{18}$$

The ZM model describes an infinite population, since $S = \int_0^B g(\pi)\, d\pi = \infty$, and its structural type distribution

$$\begin{aligned} G(\rho) &= \int_\rho^B g(\pi)\, d\pi = C \cdot \int_\rho^B \pi^{-\alpha-1}\, d\pi = C \cdot \left[\frac{\pi^{-\alpha}}{-\alpha}\right]_\rho^B \\ &= \frac{C \cdot \rho^{-\alpha}}{\alpha} - \frac{C \cdot B^{-\alpha}}{\alpha} = \frac{C/\alpha}{\rho^\alpha} - \frac{1-\alpha}{B \cdot \alpha} \end{aligned}$$

[7] The constraints on the parameter α follow from $0 < 1/a < 1$. C is a normalising constant and will be determined from (13). The upper cutoff point B is necessary since the model would predict types with probability $\pi > 1$ otherwise. B should roughly correspond to the probability π_1 of the most frequent type.

is identical to (15) with $a = \alpha^{-1}$ and $b = (1-\alpha)B^{-1}\alpha^{-1}$ for any values of ρ where $G(\rho) \in \mathbb{N}$. Thus, (16) can indeed be understood as a continuous extension of the Zipf-Mandelbrot law.

$$\begin{aligned} E\big[V_m\big] &= \int_0^\infty \frac{(N\pi)^m}{m!} e^{-N\pi} g(\pi)\, d\pi = \frac{C}{m!} \int_0^B (N\pi)^m e^{-N\pi} \pi^{-\alpha-1}\, d\pi \\ &= \frac{C}{m!} \int_0^{NB} t^m e^{-t} \left(\frac{t}{N}\right)^{-\alpha-1} \frac{1}{N}\, dt = \frac{C}{m!} N^\alpha \int_0^{NB} t^{m-\alpha-1} e^{-t}\, dt \\ &\approx \frac{C}{m!} N^\alpha \int_0^\infty t^{m-\alpha-1} e^{-t}\, dt \end{aligned}$$

In the second line, the substitution $t := N\pi$ has been made. The approximation in the last line is justified for $NB \gg m$ (which should always be the case for the large samples that are of interest here) where the integral $\int_{NB}^\infty t^{m-\alpha-1} e^{-t}\, dt$ is vanishingly small. Thus, $E\big[V_m\big]$ is reduced to the gamma integral $\int_0^\infty t^{m-\alpha-1} e^{-t}\, dt = \Gamma(m-\alpha)$ (Weisstein, 1999: *s.v. Gamma Function*) and we obtain the concise expression

$$E\big[V_m\big] = \frac{C}{m!} \cdot N^\alpha \cdot \Gamma(m-\alpha). \tag{19}$$

The computation of $E\big[V\big]$ involves an improper integral solved by partial integration:

$$\begin{aligned} E\big[V\big] &= \int_0^\infty (1 - e^{-N\pi}) g(\pi)\, d\pi \approx C N^\alpha \int_0^\infty (1 - e^{-t}) t^{-\alpha-1}\, dt \\ &= C N^\alpha \cdot \lim_{A \downarrow 0} \left(\int_A^\infty t^{-\alpha-1}\, dt - \int_A^\infty e^{-t} t^{-\alpha-1}\, dt \right) \\ &= C N^\alpha \cdot \lim_{A \downarrow 0} \left(\left[\frac{t^{-\alpha}}{-\alpha}\right]_A^\infty - \left[e^{-t} \frac{t^{-\alpha}}{-\alpha}\right]_A^\infty - \int_A^\infty e^{-t} \frac{t^{-\alpha}}{-\alpha}\, dt \right) \\ &= C N^\alpha \cdot \lim_{A \downarrow 0} \Bigg(\underbrace{\big(1 - e^{-A}\big) \cdot \frac{A^{-\alpha}}{\alpha}}_{= O(A^{1-\alpha}) \to 0} + \underbrace{\frac{\Gamma(1-\alpha, A)}{\alpha}}_{\to \Gamma(1-\alpha)/\alpha} \Bigg) \end{aligned}$$

where $\int_A^\infty e^{-t} t^{-\alpha}\, dt = \Gamma(1-\alpha, A)$ is the upper incomplete gamma function (Weisstein, 1999, *s.v. Incomplete Gamma Function*). This leads to

$$E\big[V\big] = C \cdot N^\alpha \cdot \frac{\Gamma(1-\alpha)}{\alpha}. \tag{20}$$

Consequences of (19) and (20) are the recurrence relation

$$\frac{E\big[V_{m+1}\big]}{E\big[V_m\big]} = \frac{\Gamma(m+1-\alpha)}{(m+1)!} \cdot \frac{m!}{\Gamma(m-\alpha)} = \frac{m-\alpha}{m+1}, \tag{21}$$

a relative frequency spectrum

$$\frac{E\big[V_m\big]}{E\big[V\big]} = \frac{\alpha \cdot \Gamma(m-\alpha)}{\Gamma(m+1) \cdot \Gamma(1-\alpha)} \tag{22}$$

which is independent of the sample size N (Baayen, 2001: 118), and a power law

$$E\big[V(N)\big] = C' \cdot N^\alpha \quad \text{with} \quad 0 < \alpha < 1 \tag{23}$$

for the vocabulary growth curve. Equation (23) is known as Herdan's law (Herdan, 1964) in quantitative linguistics and as Heaps' law (Heaps, 1978) in information retrieval.

The appeal of the ZM model lies in its mathematical elegance and numerical efficiency. Computation of the expected frequency spectrum and similar statistics is fast and accurate, using implementations of the complete and incomplete gamma function that are provided by many scientific libraries. Moreover, due to the simple form of $g(\pi)$ many other important integrals such as

$$E\big[V_{m,\rho}\big] = \int_0^\rho \frac{(N\pi)^m}{m!} e^{-N\pi} g(\pi)\, d\pi \tag{24}$$

for $0 < \rho < B$ have closed-form solutions and can be studied analytically.

4. The finite Zipf-Mandelbrot (fZM) LNRE model

Although the ZM model is theoretically well-founded as a model for random character sequences, its assumption of an infinite vocabulary is unrealistic for natural-language data. In order to achieve a better approximation of such frequency distributions, the finite ZM model introduces an additional lower cutoff point $A > 0$ for the type density:

$$g(\pi) := \begin{cases} C \cdot \pi^{-\alpha-1} & A \le \pi \le B \\ 0 & \text{otherwise} \end{cases}, \tag{25}$$

which implies that there are no types with probability $\pi < A$ in the population. The normalising constant C is determined from (13) as

$$C = \frac{1-\alpha}{B^{1-\alpha} - A^{1-\alpha}}, \tag{26}$$

and the population size is

$$S = \frac{C}{\alpha} \cdot (A^{-\alpha} - B^{-\alpha}) = \frac{1-\alpha}{\alpha} \cdot \frac{A^{-\alpha} - B^{-\alpha}}{A^{1-\alpha} - B^{1-\alpha}}. \tag{27}$$

Again, the structural type density $G(\rho)$ is identical to (15), with $G(\rho) = S$ for $\rho \le A$. The expectations of V_m and V are calculated to be

$$E\big[V_m\big] = \frac{C}{m!} \cdot N^\alpha \cdot \Gamma(m-\alpha, NA), \tag{28}$$

$$E\big[V\big] = C \cdot N^\alpha \cdot \frac{\Gamma(1-\alpha, NA)}{\alpha} + \frac{C}{\alpha \cdot A^\alpha}\left(1 - e^{-NA}\right). \tag{29}$$

There are no simple expressions for the recurrence relation (21) and the relative frequency spectrum (22). Although much of the mathematical elegance of the ZM model has been lost, the fZM model is still numerically efficient, and many integrals like (24) have closed-form expressions involving incomplete gamma functions.

5. Some other (related) LNRE models

Rouault (1978) has studied the properties of random character processes and shown that their observed relative frequency spectrum V_m/V converges to the expression (22) predicted by the ZM model for $N \to \infty$. This result provides theoretical support for its use as a model of large samples of random (or nearly random) text.

In the literature on lexical statistics, models for word frequency distributions are often based on a power law similar to (16). Sometimes, a decay factor $e^{-\lambda\pi}$ is used instead of the arbitrary cutoff point B. Such a model is introduced by Good (1953: 248) as a "Pearson Type III" distribution for $\alpha < 0$, and generalised to the range $0 < \alpha < 1$ by Engen (1974). Multiplication with a second decay factor $e^{-\mu/\pi}$ instead of the lower cutoff point A of the fZM model leads to Equation (50) of Good (1953: 249). Good refers to it as a mixture "between Pearson's Types III and V" and remarks that it is "analytically unwieldy". Sichel (1971, 1975) works out this model under the name Generalized Inverse Gauß-Poisson (GIGP), using the type density

$$g(\pi) = \frac{(2/bc)^{\gamma+1}}{2K_{\gamma+1}(b)} \pi^{\gamma-1} e^{-\frac{\pi}{c} - \frac{b^2 c}{4\pi}} \tag{30}$$

with parameters b, c, and γ (Baayen, 2001: Sec. 3.2.2.). Carroll (1967) and Holgate (1969) assume a log-normal distribution for the population frequencies of words or species in a biological sample, citing Preston (1948) for a theoretical motivation. The resulting type density is

$$g(\pi) = \frac{1}{\sigma\sqrt{2\boldsymbol{\pi}}} \frac{1}{\pi^2} e^{-\frac{1}{2\sigma^2}(\log\pi - \mu)^2} \tag{31}$$

with parameters μ and σ (Baayen, 2001: Sec. 3.2.2.).[8] (Baayen, 2001: Sec. 3.2.3) presents several other models based on variants of Zipf's law for the expected frequency spectrum at a certain Zipf sample size Z (see also Good, 1953: 249). Although the expectations of V_m and V can be computed for arbitrary sample sizes N using extrapolation techniques, none of these models can be reformulated as a population model (Baayen, 2001: 94). Implementations of the GIGP, log-normal, and several of the Zipf models are available in the `lexstats` package distributed with (Baayen, 2001). Of the various Zipf models, the Yule-Simon model (Simon, 1960) is found to be useful and numerically manageable.

6. Empirical data

In order to see how well the ZM and fZM models describe real-world data, they have been applied to nouns and adjective-noun cooccurrences extracted from the 100-million word British National Corpus (BNC) and a 225-million word corpus of German newspaper text from the 1990's (HGC). The following four data sets were used. **BNC-N**: 19 million instances of nouns extracted from the BNC corpus, and filtered with regular expressions to weed out non-words ($N = 19 \times 10^6, V = 217\,527$). **HGC-N**: 48 million instances of nouns extracted from the HGC corpus, and checked with a morphological analyser (Lezius *et al.*, 2000) ($N = 48 \times 10^6, V = 1\,556\,203$). **BNC-AN**: 4 million instances of adjacent adjective-noun pairs from the BNC corpus. Both the adjective and the noun were checked with regular expressions ($N = 4 \times 10^6, V = 1\,391\,498$). **HGC-AN**: 12 million instances of adjectives modifying nouns within a noun phrase, extracted using part-of-speech patterns. This simple extraction method has been found to reach excellent precision (Evert and Kermes, 2003). Both the adjective and the noun were validated with a morphological analyser ($N = 12 \times 10^6, V = 3\,621\,708$).

The Herdan law and the size-invariant relative frequency spectrum, which are characteristic properties of the ZM model, have repeatedly been critcised as unrealistic (e.g. Baayen, 2001: 118). Figure 1 shows the development of the relative frequency spectrum up to $m = 5$ for the HGC-AN data set (left panel). After approximately 2 million tokens, the relative spectrum has converged and is nearly constant afterwards. Likewise, the relative error of the Herdan law $E\big[V(N)\big] = C \cdot N^\alpha$ with $\alpha = 0.87$ (determined by linear regression) remains below 1% after

[8] The constant $\boldsymbol{\pi}$ is printed in bold font to distinguish it from the type probability π.

the first 4 million tokens (right panel). Together with similar results for the other three data sets, this is a strong indication that the ZM and fZM models may indeed be well suited for the type of frequency data represented by these data sets.

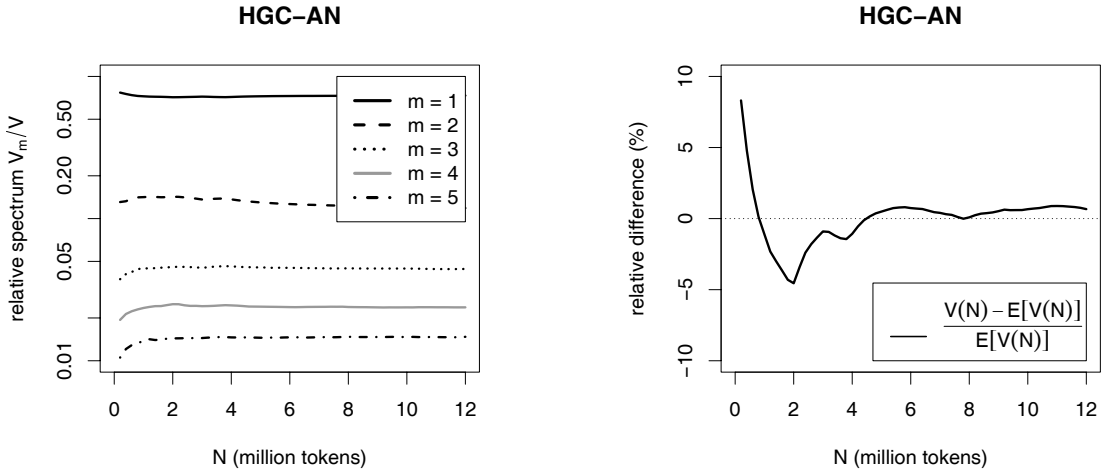


Figure 1. Development of relative frequency spectrum and relative error of Herdan law (Heaps' law) with $\alpha = 0.87$ *for the* HGC-AN *data set*

7. Evaluation of the Zipf-Mandelbrot models

Both the ZM and the fZM model were implemented using the freely available statistical computation software R,[9] and fitted to the four data sets described in Section 6. For the infinite ZM model, the parameter α can be estimated directly from (22) for $m = 1$:

$$\alpha = \frac{E\big[V_1\big]}{E\big[V\big]} \approx \frac{V_1}{V} \tag{32}$$

(see also Rouault, 1978: 172). However, Equation (32) turned out to give unsatisfactory results, so the parameters for both models were estimated through non-linear minimisation of a goodness-of-fit chi-squared statistic for the first 15 spectrum elements, with the additional constraint $E\big[V\big] = V$. Goodness-of-fit was measured with a multi-variate chi-squared test, following (Baayen, 2001: Sec. 3.3.) and using the `lexstats` implementation. The results are shown in Table 1.[10]

The fZM model gives considerably better approximations of the observed frequency spectrum than the ZM model, especially for the adjective-noun data sets where the distribution of population probabilities is much more skewed (indicated by a larger value of α). It is worth noting that the fZM model is entirely consistent with the BNC-N data set: $\chi^2_{13} = 22.20$ corresponds to a p-value of $p \approx 0.0524$ and the model is thus accepted at the 5% significance level.

A graphic representation of the accordance between the expected and observed frequency spectrum for the HGC-AN data set is shown in Figure 2. Surprisingly, the estimated lower cutoff points ($A = 9.267 \times 10^{-9}$ for BNC-AN and $A = 1.576 \times 10^{-9}$ for HGC-AN) are already quite close to the observed relative frequency of the hapax legomena ($p = 1/N$). According to the

[9] http://www.r-project.org/

[10] Note that the χ^2 statistic for the ZM model has $df = 14$ because 2 parameters were estimated from the observed spectrum. Likewise, the statistic for the fZM model with 3 estimated parameters has $df = 13$.

data set	ZM model		fZM model		
	α	χ^2_{14}	α	S	χ^2_{13}
BNC-N	0.4686416	80.75	0.4728356	4 021 728	22.20
HGC-N	0.6181580	27015.67	0.6663519	16 325 666	591.72
BNC-AN	0.7145849	313472.66	0.9168508	9 048 002	9364.46
HGC-AN	0.7441247	441448.77	0.9134667	37 983 975	1855.59

Table 1. Estimated Zipf parameter α, population size S, and goodness-of-fit statistic χ^2 for the ZM and fZM models applied to the four data sets of Section 6.

predicitions of the fZM model, increasing the sample 100-fold ($N \approx 10^9$) would already leave the LNRE zone, with all expected frequencies greater than 1 (cf. Baayen, 2001: Sec. 2.4.).

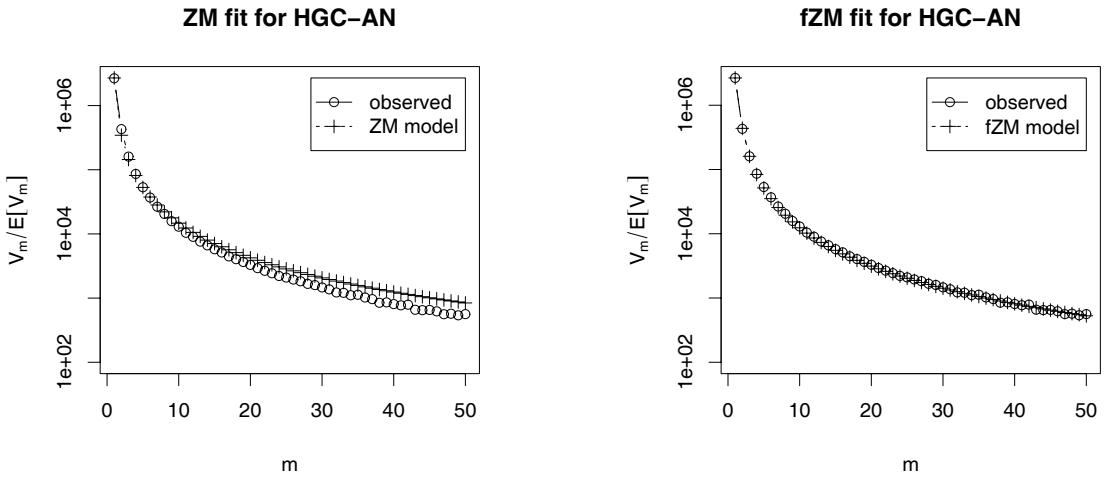


Figure 2. Expected frequency spectrum of ZM (left panel) and fZM (right panel) models compared to observed spectrum for the HGC-AN *data set (logarithmic scale, up to $m = 50$)*

A possible explanation for this counter-intuitive result is provided by term clustering effects, which violate the randomness assumption and cause the number V_1 of hapax legomena to be less than predicted by a random sample model. Such clustering effects can be detected with a dispersion test as described by Baayen (2001: Sec. 5.1). For the HGC-AN data set, a highly significant deviation from the randomness assumption ($p \approx 0$) was found with this test.[11]

For comparison, I also applied the GIGP, log-normal, and Yule-Simon models (see Section 5) to the four data sets, using the implementations included in the `lexstats` package with automatic parameter optimisation. Goodness-of-fit results for the HGC-AN data set range from $\chi^2_{14} = 259800.05$ (log-normal) to $\chi^2_{13} = 63531.61$ (Yule-Simon); for the numerically simpler GIGP model with $\gamma = -0.5$, no reasonable fit could be achieved. Results for the BNC-N data set range from $\chi^2_{13} = 36562.75$ (log-normal) to $\chi^2_{13} = 1267.81$ (GIGP).

[11] For the application of the dispersion test, the $N = 12 \times 10^6$ tokens were divided into 12 000 equally-sized chunks of 1 000 tokens each. The probability that a dis legomenon (a type with $f = 2$) is underdispersed, i.e. both its occurrences are in the same chunk, is found to be $p = 8.325 \times 10^{-5}$. Therefore, about 36 of the $V_2 = 430\,277$ dis legomena are expected to be underdispersed if the randomness assumption is correct. However, in the HGC-AN data set, there were 35 677 underdispersed dis legomena, which is a highly significant deviation from the expected value (according to a binomial test).

To conclude, the ZM model with its elegant formulation and convenient analytical properties achieves a goodness-of-fit comparable to that of the other LNRE models. The less elegant fZM model consistently outperforms its competitors and has the additional benefit of a fast and robust numerical implementation.

References

Baayen R.H. (2001). *Word Frequency Distributions*. Kluwer.

Baayen R.H. and Renouf Ant. (1996). Chronicling the Times: Productive lexical innovations in an English newspaper. *Language*, vol. (72/1): 69-96.

Carroll J.B. (1967). On sampling from a lognormal model of word frequency distribution. In Kučera H. and Francis W.N. (Eds), *Computational Analysis of Present-Day American English*: 406-424.

Engen St. (1974). On species frequency models. *Biometrika*, vol. (61/2): 263-270.

Evert St. and Kermes H. (2003). Experiments on candidate data for collocation extraction. In *Companion Volume to the Proceedings of the 10th Conference of The European Chapter of the Association for Computational Linguistics*: 83-86.

Evert St. and Krenn Br. (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*: 188-195.

Good I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, vol. (40/3-4): 237-264.

Heaps H.S. (1978). *Information Retrieval – Computational and Theoretical Aspects*. Academic Press.

Herdan G. (1964). *Quantitative Linguistics*. Buttersworths.

Holgate P. (1969). Species frequency distributions. *Biometrika*, vol. (56/3): 651-660.

Khmaladze E.V. (1987). *The statistical analysis of large number of rare events*. Technical Report MS-R8804. Department of Mathematical Statistics, CWI, Amsterdam, Netherlands.

Lezius W., Dipper St. and Fitschen A. (2000). IMSLex – representing morphological and syntactical information in a relational database. In Heid U., Evert S., Lehmann E. and Rohrer C. (Eds), *Proceedings of the 9th EURALEX International Congress*: 133-139.

Li W. (1992). Random texts exhibit zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, vol. (38/6): 1842-1845.

Mandelbrot B. (1962). On the theory of word frequencies and on related Markovian models of discourse. In Jakobson R. (Ed.), *Structure of Language and its Mathematical Aspects*: 190-219.

McNeil D.R. (1973). Estimating an author's vocabulary. *Journal of the American Statistical Association*, vol. (68): 92-96.

Miller G.A. (1957). Some effects of intermittent silence. *The American Journal of Psychology*, vol. (52): 311-314.

Powers D.M.W. (1998). Applications and explanations of Zipf's law. In Powers D.M.W. (Ed.), *Proceedings of New Methods in Language Processing and Computational Natural Language Learning*: 151-160.

Preston F.W. (1948). The commonness, and rarity, of species. *Ecology*, vol. (29): 254-283.

Quan Ha L., Sicilia-Garcia E.I., Ming J. and Smith F.J. (2002). Extension of Zipf's law to words and phrases. In *Proceedings of COLING 2002*.

Rouault A. (1978). Lois de Zipf et sources markoviennes. *Annales de l'Institut H. Poincaré (B)*, vol. (14): 169-188.

Sichel H.S. (1971). On a family of discrete distributions particularly suited to represent long-tailed frequency data. In Laubscher N. F. (Ed.), *Proceedings of the Third Symposium on Mathematical Statistics*: 51-97.

Sichel H.S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, vol. (70): 542-547.

Simon H.A. (1960). Some further notes on a class of skew distribution functions. *Information and*

Control, vol. (3): 80-88.

Weisstein E.W. (1999). *Eric Weisstein's World of Mathematics*. Wolfram Inc. On-line resource at http://mathworld.wolfram.com/

Zipf G.K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

Quantité d'information échangée : une nouvelle mesure de la similarité des mots

Cédrick Fairon[1], Ngoc-Diep Ho[2]

[1]UCL – FLTR – CENTAL – 1348 Louvain-la-Neuve – Belgique
[2]UCL – FSA – INMA – 1348 Louvain-la-Neuve – Belgique
fairon@tedm.ucl.ac.be, ho@inma.ucl.ac.be

Abstract

There are a lot of methods for measuring semantic similarity between words that are based on different approaches. This document proposes a method based on the analyses of a dictionary. The definitions of words in the dictionary create a network. Its nodes are the headwords found in the dictionary and its edges represent relations between a headword and the words found in its definitions. The meaning of a word is defined by the *total quantity of information*, which each element of its definition contributes in. The similarity between two words is defined by the maximal *quantity of information exchanged* between them through the network.
In order to assess the performance, our measure of similarity will be compared with others measures and some applications based on our similarity will be also constructed.

Résumé

Il existe beaucoup de méthodes pour mesurer la similarité entre mots et ces méthodes se basent souvent sur des approches différentes. La recherche que nous présentons a pour but de proposer une nouvelle méthode basée sur l'analyse d'un dictionnaire. Les définitions du dictionnaire créent un réseau dont les nœuds sont les entrées lexicales du dictionnaire, les arcs sont des liens représentants la relation entre une entrée et les mots de ses définitions. Le sens d'un mot dépend de la *quantité totale* d'*information* que chaque mot dans sa définition va lui communiquer. La similarité entre 2 mots est définie par la *Quantité d'Information Echangée* (QIE) entre 2 mots, à travers le réseau.
Notre mesure de similarité sera comparée avec d'autres mesures et quelques applications basées sur cette mesure seront réalisées.

Mots-clés : similarité de mots, extraction de synonymes, filtre sémantique, flot maximal.

1. Introduction

Dans les dictionnaires explicatifs comme le *Petit Robert*, on trouve très souvent des synonymes ou des antonymes pour un mot quelconque. Par exemple, le mot « *maison* » et le mot « *logement* » sont synonymes, le mot « *maison* » et le mot « *abri* » sont aussi synonymes. Mais, comment peut-on dire que la connexion entre « *maison* » et « *logement* » est plus forte que celle entre « *maison* » et « *abri* » ? La réponse à cette question implique la notion de *similarité des mots* qui peut se représenter par une valeur scalaire qui définit comment 2 mots se relient. Plus concrètement, si la similarité entre le mot m_1 et le mot m_2 est quantifiée par $sim(m_1, m_2)$, on peut dire que « *maison* » est plus proche de « *logement* » que de « *abri* » si on a $sim(maison, logement) > sim(maison, abri)$ et vice versa.

La formalisation et la quantification de la similarité des mots ont été introduites depuis très longtemps. Cela remonte au moins à l'époque d'Aristote (384 – 322 B.C) (Budanisky, 1999), mais ces préoccupations n'avaient pas, jusqu'à il y a peu, trouvé beaucoup d'applications concrètes.

Dans le domaine du Traitement Automatique du Langage Naturel (TALN), les relations sémantiques comme la synonymie, l'antonymie, l'hyponymie, la méronymie, etc. sont des notions particulièrement importantes. Avec le développement de l'informatique et du Web, il y a chaque jour plus d'information disponible sur les pages du Web et sur les archives électroniques. C'est un avantage remarquable que l'on n'aurait jamais imaginé au début du siècle précédent. Cependant, pour les informaticiens, ça pose un problème pratique : comment trouver les informations utiles et gérer ces informations ? C'est la question qui a mené au développement du domaine de la *Recherche d'Information (RI)*. Jusqu'à présent, la plupart des méthodes utilisées dans les moteurs de recherche sont basées sur l'analyse statistique des occurrences des mots dans les documents. Cela marche bien dans beaucoup de cas, mais il y a encore des cas où ces méthodes ne sont pas satisfaisantes. Par exemple, quand on lance une recherche avec le mot clé : « *oiseau* », on ne veut pas seulement obtenir les documents contenant le mot « *oiseau* ». On attend aussi les documents qui contiennent les synonymes ou les mots étroitement liés avec le mot « *oiseau* » au niveau sémantique. Dans ce contexte, les *mesures de similarité* entre les mots sont particulièrement utiles.

Penchons-nous un instant sur la structure d'un dictionnaire explicatif. On peut constater en observant une entrée lexicale donnée que les mots trouvés dans sa définition jouent le rôle de *fournisseur* d'information. On remarque également que la plupart de ces mots seront définis par ailleurs dans d'autres notices où ils occuperont la position d'entrée lexicale et où ils seront cette fois en position de *récepteur* d'information. Dans ce contexte, la quantité d'information qu'un mot peut recevoir et fournir dépend donc de chaque mot et peut être calculée avec l'aide de la théorie de l'information. C'est sur cette constatation que nous nous fondons pour définir notre mesure de similarité basée sur la *Quantité d'Information Echangée* (QIE) et l'appliquer à *l'extraction automatique de synonymes* (cf. section 5) et au *filtrage sémantique* (cf. section 6).

2. Travaux antérieurs

Plusieurs méthodes ont été proposées ces dernières années pour mesurer la similarité entre des mots et développer des applications sémantiques dans le domaine du traitement automatique du langage. Elles peuvent être classifiées dans 4 catégories :

- celles qui exploitent un dictionnaire explicatif (Kozimo et Furugori, 1993 ; Kozima et Ito, 1995 ; …) ;
- celles qui exploitent Wordnet (Hisrt et St-Onge, 1997 ; Leacock et Chodorow, 1998 ;…) ;
- celles qui exploitent Wordnet et un corpus (Jiang et Conrath, 1997 ; Resnik, 1995 ; Lin, 1998c ;…) ;
- celles qui exploitent un thésaurus (Okumura et Honda, 1994 ;…).

Notre méthode se situe dans la première catégorie.

3. Similarité basée sur la Quantité d'Information Echangée (QIE)

La nouvelle définition de similarité que nous proposons se base sur des réseaux d'interconnexion entre des concepts et le contenu informationnel de ces concepts. Dans un réseau de ce type, les concepts jouent le rôle des nœuds et les rapports entre ces concepts sont représentés par les arcs. Nous expliquerons plus loin comment nous construisons un tel réseau. L'information contenue par les concepts et celle des rapports entre concepts seront consi-

dérées seulement au niveau quantitatif (i.e. la quantité d'information) mais pas au niveau qualitatif (i.e. la sémantique du type d'information). Ces idées initiales nous offrent une intuition très importante pour la définition de notre mesure de similarité.

Intuition : la similarité entre le concept A et le concept B dépend de l'information que A peut transférer vers B et de l'information que B peut transférer vers A. Autrement dit, la similarité de A et B dépend de la *Quantité d'Information Echangée (QIE)* entre A et B.

La supposition qui mène à une nouvelle définition de la similarité est la suivante :

Supposition : la description d'un concept A est constituée par la quantité d'information que ses objets voisins lui transfèrent.

Considérons un exemple avec un concept A qui est en rapport avec n autres concepts O_1, O_2, …, O_n.

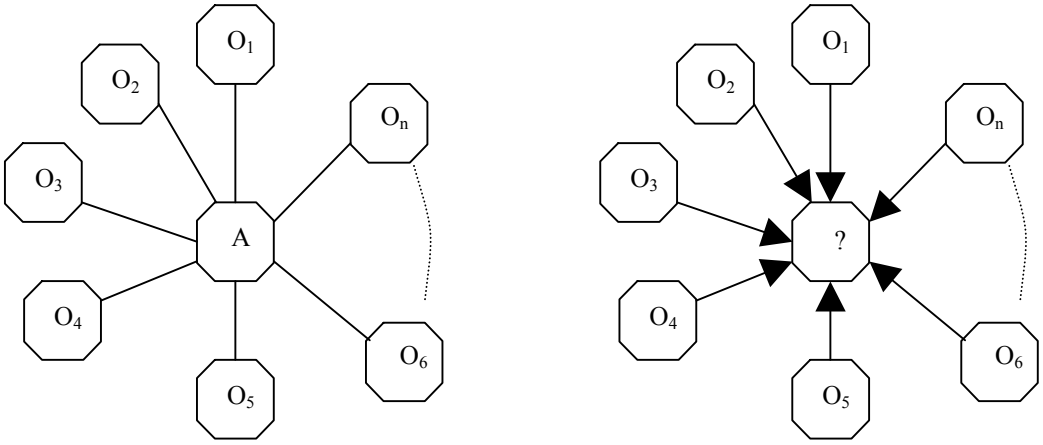


Figure 1. Un objet avec ses voisins

Dans la figure 1 à gauche, le concept A se trouve au milieu de ses voisins. Supposons que l'on ne connaisse pas A (cf. figure 1 à droite). Dans un tel cas, on ne connaît pas A, mais l'incertitude est réduite grâce à l'existence de O_1, O_2, …, O_n.

Et dans la théorie de l'information, l'information est toujours une mesure de la réduction de l'incertitude. Donc, on peut dire que les concepts O_1, O_2, …, O_n ont transféré une certaine quantité d'information pour réduire l'incertitude de A.

Dans un domaine où il existe un modèle probabiliste, et grâce à la théorie de l'information, le contenu d'information d'un concept A est calculé par la formule logarithmique de I :

$$I(A) = -\log(P(A)) \qquad (3.1)$$

où $P(A)$ est la probabilité de A. En partant de notre supposition, nous pouvons dire que la somme d'information que A reçoit de ses voisins est $I(A)$. On ne peut pas déterminer la quantité d'information qu'un O_i transfère vers A sur le lien de O_i à A, parce que cela dépend de la similarité entre des objets et que celle-ci n'a pas encore été définie. Mais on peut faire la remarque suivante :

Remarques : Un concept qui a plus d'information (faible probabilité d'occurrence) peut donner plus d'information.

Donc, on peut estimer la quantité d'information que chaque O_i peut transférer vers A sur le lien de O_i à A d'une manière très simple :

(Information transférée de O_i vers A) *(3.2)*

où $w_i = \frac{I(O_i)}{\sum_j I(O_j)}$

Maintenant, on va formaliser la similarité à partir des liens d'information que l'on vient de construire pour capturer notre intuition.

Considérons le cas où on a 2 concepts A et B dans un réseau d'interconnexion des concepts Ω.

- Le concept B peut transférer son information vers le concept A via un lien direct (B est un voisin de A) ou/et via des liens indirects (en passant par d'autres objets).
- Il existe au moins un chemin qui mène de A à B et vice versa.
- Les chemins qui connectent 2 concepts A et B décrivent la relation entre ces 2 concepts.

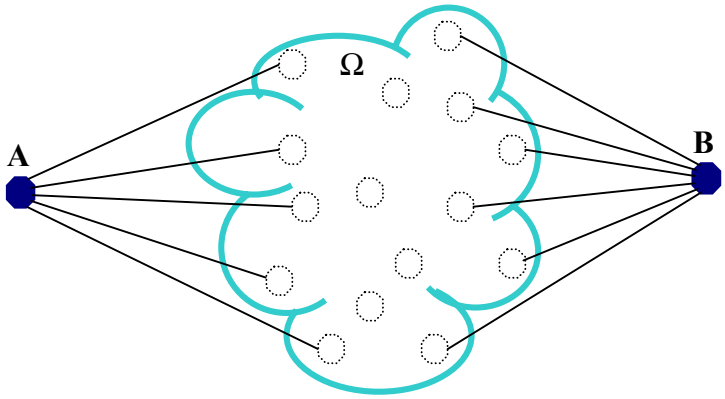


Figure 2. Deux concepts dans un réseau d'interconnexion

La quantité d'information que le concept B peut vraiment transférer vers le concept A via le réseau Ω est calculée par le *flot maximal d'information* de B à A, notée par $fmi_\Omega(B,A)$. Donc, la similarité entre A et B (qui correspond à la quantité d'information échangée entre A et B) est décrite par la formule :

$$sim_{QIE}(A,B) = f(mfi_\Omega(A,B), mfi_\Omega(B,A)) \quad (3.3)$$

Le choix du Ω dépend du domaine d'application. Les choix les plus naturels pour la fonction f sont la moyenne et la moyenne exponentielle. Et ça fournit les 2 formules suivantes :

$$sim_{QIE1}(A,B) = \frac{mfi_\Omega(A,B) + mfi_\Omega(B,A)}{2} \quad (3.4)$$

et

$$sim_{QIE2}(A,B) = \sqrt{mfi_\Omega(A,B).mfi_\Omega(B,A)} \quad (3.5)$$

Nous appelons cette mesure la similarité basée sur la *Quantité d'Information Echangée (QIE)*. Dans la section suivante, nous allons brièvement expliquer comment calculer la similarité entre des mots anglais. Ensuite, nous présenterons 2 applications de cette mesure de similarité : pour l'extraction de synonymes et pour le filtrage sémantique.

4. Similarité QIE pour les mots anglais

4.1. *Dictionnaire Webster pour l'anglais*

Grâce au projet Gutenberg, une version de dictionnaire 1913 US Webster (Webster) est publiée sur le Web dans 27 fichiers HTML. Chacun des 26 fichiers contient des mots commençant par le même caractère (A-Z), le dernier fichier ne contient que les mots nouveaux qui ne sont pas définis dans les 26 premiers fichiers.

Dans son rapport de stage à l'Université catholique de Louvain, P. Senellart (2001) décrit son travail de conversion du dictionnaire Webter en graphe et il analyse les propriétés de ce graphe. En lisant 27 fichiers HTML, il crée pour chaque mot du dictionnaire un nœud dans le graphe. Chaque arc (i,j) dans ce graphe représente l'occurrence de mot j dans la définition de mot i. Le graphe final comprend 112 169 nœuds et 1 398 424 arcs.

4.2. *Calcul de la similarité*

En ajoutant la capacité des liens au graphe Webster, nous avons un réseau de concept qui peut être utilisé pour calculer la similarité. En effet, nos applications utilisent rarement ce réseau complet pour calculer la similarité. Il suffit d'utiliser seulement un sous réseau qui se déduit avec tous les voisins de 2 mots dont nous devons calculer la similarité. Cette simplification nous permet d'accélérer la vitesse de calcul et d'éviter les connexions inutiles (trop longues connexions) entre les 2 mots.

4.3. *Évaluation de la similarité*

Comment peut-on juger si une mesure de similarité est performante et comment peut-on comparer deux mesures ? Ce sont des questions qui n'ont pas encore une réponse totalement convaincante.

La méthode qui nous semble la plus acceptable consiste à comparer les résultats avec les jugements humains. Dans le rapport Budanisky, 1999, 2 jeux de test ont été utilisés pour comparer plusieurs mesures de la similarité. Le premier jeu créé par Rubenstein et Goodenough (1965) contient 65 paires de mots. Le deuxième jeu créé par Millers et Charles (1991) contient 30 paires de mots. Toutes les paires dans le jeu de Millers et Charles se retrouvent dans le jeu de test de Rubenstein et Goodenough. Chaque paire de mots est associée avec un nombre qui indique la similarité jugée par des sujets humains. Pour évaluer les performances de la mesure computationnelle, le coefficient de corrélation entre les résultats de cette mesure et les jugements humains est calculé.

Le tableau 1 compare les résultats de notre méthode avec ceux d'autres méthodes. Les résultats numériques de ces mesures (Hirst-St-Onge, Jiang et Conrath, Leacock et Chodorow, Lin et Resnik) sont extraits de Budaniksy (1999).

Les résultats montrent que la méthode QIE peut donner des résultats équivalents aux autres méthodes en n'utilisant qu'un dictionnaire explicatif (en l'occurrence le Webster), tandis que d'autres méthodes utilisent des ressources plus structurées comme WordNet. Il faut mentionner en outre que nos résultats sont le fruit d'une première expérience et qu'ils pourraient être améliorés grâce à des traitements supplémentaires du dictionnaire Webster (reconnaissance des mots composés, lemmatisation, etc.). On peut donc considérer les résultats obtenus par notre mesure de similarité des mots comme prometteurs.

Méthodes	Rubenstein-Goodenough	Miller - Charles
Hirst et St-Onge	0.7861440344	0.7443990930
Jiang et Conrath	0.7812746298	0.8500267204
Leacock et Chodorow	0.8382296528	0.8157413049
Lin	0.8193023545	0.8291711020
Resnik	0.7786845861	0.7736382148
QIE1 (*f* = moyenne)	**0.7569047994**	**0.7515805974**
QIE2 (*f* = *moyenne exp.*)	**0.7859960606**	**0.8327221986**

Tableau 1. Coefficients de corrélation avec les jugements humains.

Dans les 2 sections qui suivent se trouvent deux applications qui utilisent cette similarité pour extraire des synonymes de mots anglais et pour filtrer les matériaux textuels.

5. Extraction de synonymes

L'extraction de synonymes est une application très utile. S'il est vrai que l'on peut facilement trouver des synonymes dans un thésaurus (comme *Roger*, par exemple) il n'en reste pas moins que de telles ressources n'existent pas dans toutes les langues, et que le nombre de synonymes proposés pour chaque mot est souvent limité. Par contre, un bon extracteur automatique de synonymes peut servir pour la plupart des langues et donner une longue liste de mots apparentés qui sont triés en ordre croissant ou décroissant de similarité.

La notion de « *synonymes* » dans ce sens concerne les mots liés entre eux et ayant un certain degré de similarité. Cette définition « large » convient également pour décrire les synonymes des thésaurus papier, car les synonymes que l'on y trouve ne peuvent pas être utilisés de manière interchangeable dans tous les cas.

Pour extraire les synonymes du mot *m*, nous calculons la similarité entre *m* et chaque mot dans le graphe des voisins de *m*. Ensuite, nous trions ces mots en fonction de ces similarités et proposons les *n* premiers mots comme synonymes de *m*.

Dans les tableaux 2 et 3, nous comparons les synonymes donnés par notre méthode avec ceux donnés par les méthodes présentées dans Senellart (2001) pour le mot *disappear* et le mot *sugar*.

Les synonymes de tous les mots du Webster ont déjà été extraits et peuvent être consultés grâce à une interface Web à l'adresse : http://cental.fltr.ucl.ac.be/synonyms.

6. Filtre sémantique

Le fonctionnement d'un filtre sémantique est caractérisé par l'élimination (ou la sélection) des éléments linguistiques (mot, phrase, texte, etc.) appartenant à une série de domaines thématiques spécifiques. Un domaine peut être décrit par une liste créé à la main de mots fortement liés à ce domaine (par exemple : finance={bilan, capital, capitaux, banque, bourse, compte, actions, opa, etc.}).

	Distance	**Senellart**	**ArcRank**	**QIE_L**	**WordNet**
1	Vanish	Vanish	Epidemic	Vanish	Vanish
2	Pass	Pass	Dissapearing	Fade	go away
3	Wear	Die	Port	Wear	End
4	Die	Wear	Dissipate	Die	Finish
5	Light	Faint	Cease	Pass	Terminate
6	Fade	Fade	Eat	Dissipate	Cease
7	Faint	Sail	Gradually	Faint	
8	Port	Light	Instrumental	Light	
9	Absorb	Dissipate	Darkness	Evanesce	
10	Dissipate	Cease	Efface	Disappearing	

Tableau 2. Synonymes de Disappear

	Distance	**Senellart**	**ArcRank**	**QIE_L**	**WordNet**
1	Cane	Cane	Granulation	Inversion	Sweetening
2	Starch	Starch	Shrub	Dextrose	Sweentener
3	Juice	Sucrose	Sucrose	Sucrose	Carbonhydrate
4	Obtained	Milk	Preserve	Lactose	Saccharide
5	Milk	Sweet	Honeyed	Cane	organic compound
6	Sucrose	Dextrose	Property	Sorghum	Saccarify
7	Molasses	Molasses	Sorghum	Candy	Sweeten
8	Sweet	Juice	Grocer	Grain	Dulcify
9	White	Glucose	Acetate	Root	Edulcorate
10	Plants	Lactose	Saccharine	Starch	Dulcorate

Tableau 3. Synonymes de Sugar

Le filtre le plus simple que l'on puisse imaginer est donc un programme qui éliminerait ou sélectionnerait des textes contenant au moins une occurrence de l'un des mots clés du domaine traité. Une approche naïve de ce type pose trois problèmes :

- Il est difficile d'établir ces listes de mots clés : comment faire, combien en faut-il ?
- L'absence de mot clé ne garantit pas qu'un texte n'appartient pas au domaine filtré.
- La présence d'un mot clé ne garantit pas que le texte appartient au domaine filtré (ambiguïtés, métaphores, etc.).

Nous avons essayé de trouver une solution qui exploite la notion de similarité des mots que nous venons d'exposer ci-dessus. Comment donc éliminer des textes appartenant à un domaine **D** en utilisant la notion de similarité des mots ? D'abord, pour chaque texte **A**, on va calculer la similarité entre chacun de ses mots et chacun des mots qui décrit le domaine **D**. Ensuite, la similarité entre le texte **A** et le domaine **D** est calculée en combinant toutes ces similarités. Cette valeur de similarité texte-domaine permet de créer les filtres plus flexibles que l'on peut paramétrer en déterminant un seuil d'acceptation (tous les textes qui reçoivent une valeur supérieure à ce seuil sont filtrés).

Une première version de filtre thématique a été expérimentée sur 1242 phrases extraites du journal américain en ligne *Detroit Free Press*. Nous avons défini le domaine à filtrer à l'aide des mots clés suivants : *crime, crimes, criminal, criminals, victim, victims, bail, bailed, bails, jail, jails, prison, prisons, custody, bribe, bribes, bribery, fraud, frauds, theft, forgery, drug, drugs, hashish, junkie, junkies, murder, murders, murderous, kill, killed, killing, killings, killer, hijacker, hijackers, hijack, hijacked, hijacking, kidnapping, kidnap, kidnapped, kidnapper, dying, homicide, homicides, suicide, suicides, terrorist, terrorists, hostages, hostage, prisoner, prisoners, genocide, genocides, atrocity, atrocities, brutal, vengeance, violent*[1].

À chacune des 1242 phrases est associée une valeur de « degré d'appartenance » qui est la somme des similarités (au sens défini ci-dessus) entre chacun des mot de la phrase et chacun des mots clés du domaine. Après avoir trié la liste des phrases en fonction de ce degré d'appartenance, nous pouvons constater que :

a) les phrases qui contiennent au moins d'un mot clé se situent principalement en tête de la liste et toutes figurent dans la première moitié.

Pos.	Appartenance	Mots clés	Phrase
1	139.852959	2	Two hours later, members of the city's Violent Crime Task Force saw the teenager and Denisha leaving the downtown Greyhound bus station with a man, Booth said.
2	118.085315	1	Yet the cost of housing, feeding and caring for a prison inmate is about $20,000 per year, or about $40 billion nationwide using 2002 figures, according to the Sentencing Project, a nonprofit organization in Washington, D.C., that promotes alternatives to prison.
3	115.218315	2	Since 1995, growth in the federal prison system mainly reflected more incarcerated drug offenders, accounting for nearly half of the total increase, and immigration offenders, accounting for more than 20 percent of the rise.
…	…	…	…

b) Parmi les premières phrases, on trouve des phrases qui ne contiennent pas de mot clé, mais dont le sens peut être effectivement considéré comme lié au domaine. Par exemple :

Pos.	Appartenance	Mots clés	Phrase
…	…	…	…
14	93.885475	0	One senator said 95 percent of the classified pages of a congressional report released last week into the work of intelligence agencies before the attacks of Sept.11, 2001, was kept secret only to keep from embarrassing a foreign government.
…	…	…	…
16	92.221736	0	ST.LOUIS (AP) -- A missing 23-month-old girl was found safe Sunday in Detroit with a teenage runaway more than a week after the

[1] Comme on le voit, il ne s'agit pas vraiment d'un « domaine », mais d'une liste assez hétéroclite de mots liés à des sujets ayant une connotation « violente » et étant fréquemment traités dans la presse.

			toddler was abducted from her St.Louis home, authorities said.
…	…	…	…
19	90.794144	0	In a hearing open to the public, government attorneys asked Haddad about Global Relief's links with Sheikh Abdallah Azzam, a man the government says cofounded Makhtab Al-Khidemat, a precursor to Al Qaeda, and was a mentor to Osama bin Laden.
…	…	…	…
26	84.262428	0	Hudson, 68, Gilbert, 55, and Platte, 66, were convicted in April of obstructing the national defense and damaging government property last fall after cutting a fence and walking onto a Minuteman III silo site in Colorado, swinging hammers and using their blood to paint a cross on the structure.
…	…	…	…

c) Les phrases qui se trouvent à la fin de la liste ne sont pas liées au domaine.

Pos.	**Appartenance**	**Mots clés**	**Phrase**
…	…	…	…
900	4.581682	0	Do your beard and your street rod suggest you've just left the set of a ZZ Top video?
901	4.533010	0	The group, with eight representatives and nine senators, has a Republican majority.
902	4.501355	0	Ultimately, Rochester Road also will be resurfaced between Gunn and Lakeville roads, but not widened.
903	4.476029	0	But they share space in a downtown Detroit office building and occasionally wind up on the same case.
…	…	…	…
1078	1.774052	0	"What's at stake is how good will the bill be?"
1079	1.767393	0	As well they might.
1080	1.751202	0	People are now safer, stronger
1081	1.720184	0	Night Pick 4 Numbers: two, eight, nine, one.
1082	1.711899	0	"Come and look at it," Josh said.
1083	1.710149	0	"But then again, we're excited to go see my husband.
1084	1.705655	0	Jacob Hoogendyk, R-Portage, was among lawmakers who said the subsidy should remain at $5.7 million.
1085	1.702224	0	Students at Livonia's Schoolcraft College will foot a 7 percent tuition increase.
…	…	…	…
1240	0.389159	0	I couldn't eat.
1241	0.387589	0	Haju Sunim Lundquist of the Zen Buddhist Temple in Ann Arbor.
1242	0.382479	0	There are lane closures on I-94 between Wayne and I-275.

Le programme fournit une liste ordonnée en fonction de la similarité par rapport au domaine filtré, ce qui représente un avantage par rapport au filtre naïf que nous avons évoqué au début

de cette section. La liste sera donc filtrée grâce à un seuil que l'on pourra déterminer en fonction des applications en jeu. Naturellement, plus ce seuil sera haut, plus grande sera la probabilité de sélectionner des éléments qui sont liés au domaine analysé.

Nous rendons compte dans cet article des premiers tests effectués avec ce filtre et les résultats ne sont que partiels, car les développements sont toujours en cours.

Bien entendu, en parcourant la liste, nous rencontrons également des exemples montrant que les performances sont encore loin d'être idéales. Le tableau suivant montre 2 exemples qui attestent des difficultés rencontrées.

Pos.	Appartenance	Mots clés	Phrase
…	…	…	…
4	112.147987	0	On Saturday, Terry Browning Jr., beat the previous top speed of 73.734 m.p.h by 0.364 m.p.h with the 1-liter modified boat the Legend from Virginia Beach, Va.. Tommy Thompson of Cambridge, Md, answered with 75.684 m.p.h on Sunday in For Sale.
…	…	…	…
909	4.316700	0	"If there had been any information they could connect him to terrorism, why would they remove him?"
…	…	…	…

Une des difficultés rencontrées pendant cet expérience est que le filtre est utilisé pour filtrer des phrases ayant une longueur très limitée (< 40 mots/phrase), donc, il n'y a parfois pas assez d'informations pour le filtrage. Les résultats devraient sans doute être meilleurs si le filtre était appliqué à des éléments linguistiques plus longs comme : des paragraphes, des articles, etc. En outre, plusieurs démarches devraient permettre d'améliorer les résultats :

- Utilisation d'un dictionnaire plus récent. Le Webster est un ancien dictionnaire dont le lexique est loin de refléter l'état du lexique de la presse d'aujourd'hui.
- Au lieu de la somme, trouver une meilleure combinaison des similarités entre les mots des phrases et du domaine.
- Tenir compte des mots composés.
- Amélioration de la mesure de similarité elle-même.

7. Conclusion

Dans cet article, nous avons introduit une nouvelle mesure de similarité basée sur la *Quantité d'Information Echangée (QIE)* dans des réseaux de concepts. La nouvelle mesure a été implémentée pour calculer la similarité entre mots (en anglais) en utilisant le dictionnaire *Webster 1913*. La comparaison numérique que nous avons réalisée montre que la précision de notre méthode est équivalente à celle d'autres méthodes existantes et même si ces méthodes utilisent de «meilleures» ressources comme par exemple WordNet. Par ailleurs, notre approche est facilement adaptable à la plupart des langues, car elle nécessite peu de ressources (un simple dictionnaire explicatif). Par conséquent, notre méthode peut être utilisée dans des applications qui doivent être portables dans plusieurs langues.

Notre extracteur de synonymes et le filtre sémantique sont des exemples concrets d'applications pouvant être développées sur ces bases. La qualité des résultats obtenus nous

permet de penser également à beaucoup d'autres applications comme : la recherche d'un mots à partir de son explication (un animal qui vole → c'est un oiseau), la mise en correspondance d'ontologies, etc.

Références

Budanisky A. (1999). Lexical Semantic Relatedness and Its Applications in Natural Language Processing. *Rapport Technique* CSRG-390, *Computer Research Group* – Université de Toronto.

Budanisky A. et Hirst G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources, Second meeting of the North Americasimn Chapter of the Association for Computational Linguistics,* Pittsburgh.

Ford L.R. et Fulkerson D.R. (1962). *Flows in Networks*. Princeton Univ. Press.

Hirst G. et St-Onge D. (1997). Lexical chains as representations of context for the detection and correction of malapropisms. In Fellbaum Chr. (Ed.), *WordNet: An electronic lexical database*. MIT Press.

Jannink J. et Wiederhold G. (1999).thesauruss Entry Extraction from an On-line Dictionary. In *Proceedings of Fusion '99*, Sunnyvale CA.

Jiang J. et Conrath D.W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of the 10th International Conference: Research on Computational Linguistics (ROCLING X)*, Academica Sinica : 19-33.

Kozima H. et Furugori T. (1993). Similarity between words computed by spreding activation on an English Dictionary. In *Proceedings of EACL-93 (Utrecht)* : 232-239.

Kozima H. et Ito A. (1995). Context-Sensitive Word Distance by Adaptive Scaling of a Semantic Space. In Mitkov R. et Nicolov (Eds), *Recent Advances in Natural Language Processing* (une serie de *"Contemporary Issues in Linguistic Theory" 136*). John Benjamins : 111-124.

Leacock Cl. et Chodorow M. (1998). Combining Local Context and WordNet Similarity for Word Sense Identification. In Fellbaum Chr. (Ed.). *WordNet: an electronic lexical database.* MIT Press : 265-283.

Lin D. (1998b). Automatic Retrieval and Clustering of Similar Words. In *Proceedings of COLING-ACL98*, Montréal.

Lin. D. (1998c). An Information-Theoretic Definition of Similarity. In *Proceedings of International Conference on Machine Learning*, Madison, Wisconsin.

Miller G.A. et Charles W.G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes,* vol. (6/1) : 1-28.

Okumura M. et Honda T. (1994). Word Sense Disambiguation and Text Segmentation Based on Lexical Cohesion. In *Proceedings of Fifteenth International Conference on Computational Linguistics* (COLINGS-94), vol. (2) : 755-761.

Page L., Brin S., Motwani R. et Winograd T. (1998). The pagerank citation ranking : Bringing order to the Web. *Rapport Technique* Computer Science Department, Université de Stanford.

Resnik P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*.

Rubenstein H. et Goodenough J. (1965). Contextual correlates of synonymy. *CACM,* vol. (8/10) : 627--633.

Senellart P. (2001). *Extraction of information in large graphs – Automatic Search of Synonymes.* Rapport de stage. Université Catholique de Louvain.

Webster (2000). The Online Plain Text English Dictionary, http://msowww.anu.edu.au/~ralph/OPTED/, *dans le projet de Gutenberg*.

Classification et catégorisation automatiques : application à l'analyse thématique des données textuelles

Dominic Forest, Jean-Guy Meunier

UQÀM – LANCI – C.P. 8888, Succ. Centre-Ville – Montréal – Québec – Canada, H3C 3P8
forest.dominic@courrier.uqam.ca, meunier.jean-guy@uqam.ca

Abstract

Since a few years, several literature and humanities research projects have tried to integrate automatic data-processing dimensions into their objectives. In spite of each project's specificities, most of the projects' objectives concern the comprehension and the automatic processing of thematic analysis of textual data. In this paper, we present a data processing sequence adapted to thematic analysis of textual data. The specificity of this data processing sequence lies in its use of data classification and automatic categorization techniques. We present results of an experiment on a philosophical corpus.

Résumé

Depuis quelques années, plusieurs projets de recherche dans les domaines des sciences humaines et des lettres ont tenté d'intégrer des dimensions informatiques à leurs objectifs. Malgré les spécificités propres à chacun de ces projets recherche, on constate qu'un axe de recherche partagé par l'ensemble des disciplines sensibles à l'analyse de textes par ordinateur relève de la compréhension et de l'informatisation du processus d'analyse thématique des données textuelles. Dans cet article, nous présentons une chaîne de traitement adaptée à l'analyse thématique des données textuelles. La spécificité de cette chaîne de traitement réside dans son utilisation de techniques de classification et de catégorisation automatiques des données textuelles. Nous présentons les résultats d'une expérimentation sur un corpus de textes philosophiques.

Mots-clés : catégorisation, classification, analyse thématique, lecture et analyse de textes assistées par ordinateur.

1. Introduction

Depuis environ trente ans, mais surtout durant les dix dernières années, plusieurs recherches menées dans les domaines des sciences humaines et des lettres ont tenté d'intégrer des dimensions informatiques à leurs objectifs. Grâce à ces efforts d'intégration technologique, les sciences humaines ont su développer plusieurs méthodologies et applications d'analyse de textes assistées par ordinateurs. Parmi les types d'applications les plus fréquemment cités, on trouve entre autres ceux portant sur l'analyse qualitative et quantitative des données (Alexa et Zuell, 1999a et 1999b), sur l'analyse de contenu assistée par ordinateur et, de manière plus générale, sur l'analyse des données textuelles et sur la Lecture et l'Analyse de Textes Assistées par Ordinateur (LATAO) (Meunier, 1997 ; Popping, 2000).

D'autre part, le domaine des lettres et de la littérature a, lui aussi, été le lieu de réflexions théoriques et d'efforts pratiques visant à permettre l'émergence d'applications informatiques adaptées aux études littéraires (Rastier *et al.*, 1995 ; Bernard, 1999 ; Kastberg Sjöblom et Brunet, 2000). Ainsi, la rencontre entre les littéraires et les informaticiens a été le lieu d'émergence d'un nouvel axe de recherche dont les principales manifestations, inspirées de travaux en statistique et en mathématique, ont pris la forme de méthodologies d'analyse et de

logiciels destinés spécifiquement à l'analyse de textes littéraires (Bernard, 1999 ; Hockey, 2000).

Malgré les spécificités propres à chacun de ces projets recherche, on constate qu'un axe de recherche partagé par l'ensemble des disciplines sensibles à l'analyse de textes par ordinateur (tant dans le domaine des sciences humaines que des lettres) relève de la compréhension et de l'informatisation du processus d'analyse thématique des données textuelles (Louwerse et van Peer, 2002). En effet, comme le souligne Popping (2000), « in case one is interested in concept occurrences, one might use thematic analysis. This is the kind of analysis that is still applied most today » (p. ix). Ainsi, nombreux sont les chercheurs qui, à l'instar de Popping, n'hésitent pas à exprimer l'importance présentement attribuée à l'analyse thématique des documents. Pettersson (2002 : 237) est, à cet égard, des plus explicites: « [Thematics is] one of the most rapidly expanding fields in the arts, and literary studies in particular. »

2. La multidisciplinarité de la problématique de l'analyse thématique

Bien que la problématique de l'analyse thématique puise ses racines dans les travaux de Platon et d'Aristote (Sollors, 1993), plusieurs des principales recherches concernant cette problématique furent réalisées au cours du 20^{e} siècle. Parmi les principales réalisations contemporaines qui ont fortement modélisé et contribué à la compréhension des thèmes (identification et organisation) figurent celles de Tomashevsky (1925) sur la consistance et la décomposition (causale et temporaire) des thèmes en motifs (ces derniers constituant les plus petits composants d'un thème), de Aarne et Thompson (1928) sur les index de motifs, de Thompson (1946) sur la classification des motifs, de Propp (1928) sur la nécessité de développer une méthode formelle et structurée de segmentation basée sur une approche multidisciplinaire, afin d'identifier les différents motifs présents dans un texte, etc. Ces travaux fondateurs constituent les assises théoriques sur lesquelles reposent les récentes contributions à l'analyse thématique, lesquelles ont aussi largement été influencées par la distinction élaborée par l'école linguistique de Prague (principalement grâce aux travaux de N. S. Troubetskoï, de R. Jakobson et de S. O. Kartsevski) entre les concepts de « thème » (l'objet du discours) et de « rhème » (l'information relative au thème). Ce qui caractérise toutefois les principaux travaux actuels portant sur la question de l'analyse thématique concerne le niveau d'analyse, la manière dont les auteurs ont fréquemment abordé la question. En effet, tant en littérature qu'en linguistique, la majorité des travaux ont traditionnellement abordé la problématique du thème à partir d'analyses phrastiques en tentant d'identifier les éléments, principalement linguistiques, permettant de comprendre l'organisation thématique des textes (Rastier *et al.*, 1995 ; Prince, 1985 ; Louwerse et van Peer, 2002).

Malgré la richesse de la perspective linguistique, les travaux de Van Dijk (1972), Kintsch et Van Dijk (1978) et Van Dijk et Kintsch (1983) ont aussi mis en relief une approche fondée sur la distinction entre les niveaux micro-structurels (reposant, de manière générale, sur l'analyse de la phrase) et macro-structurels (dont l'objet d'analyse réside dans l'ensemble du texte, pris comme un tout cohérent et structuré) du texte. Cette approche, fondée sur des travaux provenant de plusieurs disciplines des sciences humaines semble des plus fécondes. Selon cette perspective, l'analyse thématique des textes relève d'un effort d'abstraction se situant au niveau macro-structurel et est régie par quatre principes : 1) la suppression de l'information non pertinente, 2) la sélection de l'information pertinente, 3) la généralisation des propositions retenues et 4) l'intégration des propositions dans un tout structuré et cohérent (Louwerse et van Peer, 2002). Cette théorie qui a l'avantage de « contextualiser » l'identification et l'interprétation des thèmes d'un texte a, en outre, motivé l'application de la théorie

de la sémantique latente, des calculs de cooccurrences et de corrélation au domaine de l'analyse thématique afin d'assister le chercheur dans l'interprétation des textes auxquels il est confronté. La présente recherche s'inspire partiellement de ces quatre principes en tentant toutefois d'en dépasser les limites. L'objectif de cette recherche consiste à développer une hypothèse méthodologique permettant de mettre en valeur la complémentarité des approches reposant, d'une part, sur une analyse micro-sémantique de la phrase et, d'autre part, sur une analyse macro-sémantique de la textualité (Rastier *et al.*, 1994 ; Rastier, 1996 et 2001).

On constate également qu'à travers leurs diverses perspectives d'analyse, les récents travaux dans le domaine de l'analyse thématique ont clairement démontré l'importance de la variété de points de vue que peuvent apporter les différentes disciplines concernées par cette problématique. Comme le soulignent Louwerse et van Peer (2002 : 9), « it seems more likely that in many disciplines thematics has always occupied a place but we may not have recognize it as such. » Suivant Rimmon-Kenan (1985), il semble donc nécessaire d'opter pour une approche pluridisciplinaire fondée non pas exclusivement sur des considérations linguistiques, mais plutôt sur une perspective plus large permettant de tenir compte tant des phénomènes linguistiques en jeu dans le texte que des phénomènes relevant de la textualité (processus discursif, pragmatique, etc.) sur lesquels reposent l'organisation et la structure des divers thèmes d'un corpus. Comme le soutient Giora (1985) : « la notion de thème (*topic*) discursif est indépendante de la notion de thème (*topic*) phrastique ». « Therefore, instead of describing thematics as an "undisciplined discipline" we prefer describing it as an interdisciplinary discipline, working at different levels in different areas but with a unified goal. » (Louwerse et van Peer, 2002 : 9)

3. Objectif de recherche

L'objectif général de la recherche que nous présentons ici consiste à explorer certaines méthodes de classification et de catégorisation automatiques à des fins d'analyse thématique de textes théoriques. Il vise ainsi à effectuer un transfert de concepts et de méthodologies provenant, d'une part, des recherches théoriques sur l'analyse thématique et, d'autre part, des recherches appliquées sur la catégorisation des données (intelligence artificielle, apprentissage machine, forage de textes, classification, etc.) afin de développer diverses méthodologies et applications visant à assister le chercheur en sciences humaines et en littérature dans son travail d'analyse thématique des textes. Bien que cet objectif de recherche se déploie en plusieurs volets (modélisation formelle de concepts reliés au domaine de l'analyse thématique, etc.), nous nous attarderons essentiellement dans le cadre de cet article à la présentation détaillée et à l'application exploratoire d'une chaîne de traitement spécifique adaptée à l'analyse thématique des données textuelles. Cette recherche – laquelle approfondit certaines des réflexions déjà présentées antérieurement (Forest, 2002 ; Forest et Meunier, 2000) – ne vise donc pas à procéder à une validation en profondeur de cette méthodologie d'analyse, mais plutôt d'explorer le potentiel et la fécondité d'une telle démarche.

4. Méthodologie

4.1. *La plate-forme SATIM*

Au niveau informatique, la réalisation du projet présenté consiste à concevoir et à développer une chaîne de traitement (nommée Thématico) adaptée spécifiquement à l'analyse thématique des documents textuels. Ce projet repose sur la technologie de la plate-forme informatique SATIM (Système d'Analyse et de Traitement de l'Information Multidimensionnelle). Cette plate-forme consiste en une interface permettant de structurer des modules informatiques

indépendants, de les organiser, de les faire communiquer entre eux et de les appliquer à divers domaines de traitement de l'information. La plate-forme SATIM offre des outils pour explorer des modules de traitement déjà existants afin de les agencer de façon à atteindre un but particulier (en l'occurrence la découverte et l'analyse des thèmes présents dans un corpus textuel).

4.2. *La chaîne de traitement Thématico*

La chaîne de traitement Thématico constitue une opérationnalisation informatique de l'analyse thématique (l'assistance à la découverte des principaux thèmes d'un corpus, ainsi que l'exploration et la navigation entre ces derniers) permettant d'assister le chercheur dans sa tâche d'analyse thématique des textes. La chaîne de traitement respecte une architecture classique en sept étapes.

1) L'identification des unités d'information pertinentes. Cette première étape vise à extraire du corpus les unités d'information, c'est-à-dire les unités linguistiques servant d'ancrage à l'analyse. Ces unités peuvent être des mots ou des chaînes de caractères (n-grammes) (Cavnar et Trenkle, 1994) et les domaines d'information (paragraphes, chapitres, etc.) qui seront analysés ultérieurement dans la chaîne de traitement. Certaines opérations linguistiques (suppression des mots fonctionnels, lemmatisation, étiquetage morphologique et sémantique) ou statistiques sont aussi appliquées au lexique extrait.

2) La vectorisation. Suite au découpage du corpus en unités d'information et en domaines d'information, le texte est traduit en une matrice de vecteurs (modèle vectoriel) (Salton, 1989 ; Manning et Schütze, 1999) qui représente alors chaque domaine d'information par la présence ou l'absence (binaire ou floue) des unités d'information (figure 1).

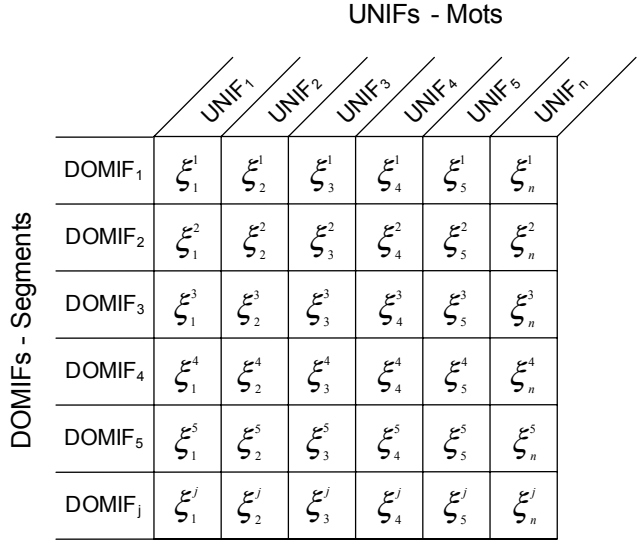


Figure 1. La matrice domaines d'information / unités d'information

3) La classification des segments. Par la suite, des classifieurs numériques sont appliqués sur la matrice. Par classifieurs numériques, nous entendons les stratégies mathématiques qui permettent la production de classes d'équivalence sur des segments. Plusieurs techniques numériques ont déjà été explorées sur le texte. Malgré certaines limites, ces approches ont présenté des résultats très positifs et se comparent avantageusement aux approches exclusivement linguistiques (Salton, 1989). Elles permettent de plus une immense économie de temps dans le parcours exploratoire d'un corpus. Technologiquement, ces approches sont incontournables

lorsqu'elles sont confrontées à de vastes corpus textuels. Dans la présente recherche, en raison de la technologie explorée et développée antérieurement par l'équipe du LANCI, nous avons privilégié comme classifieur le réseau de neurones Art (Grossberg et Carpenter, 1988).

4) La production de sous-classes lexicales différenciées. Les classifieurs produisent des regroupements (classes) de segments en raison de propriétés similaires. De ces classes, on extrait alors le lexique. Par des opérations ensemblistes simples (intersection, union, etc.), on différencie divers types de sous-classes spécifiques.

5) La catégorisation thématique. Ici, l'hypothèse est que certaines techniques de classification et de catégorisation automatiques peuvent être utilisées afin d'identifier l'organisation thématique des documents. Sur les sous-classes lexicales différenciées (étape 4) peuvent être appliquées diverses techniques de catégorisation automatique. Traditionnellement (Sebastiani, 1999 et 2003 ; Jackson et Moulinier, 2002), la catégorisation consiste à attribuer une catégorie aux domaines d'information à partir d'un ensemble prédéfini de catégories. Cette catégorisation associe une étiquette catégorielle à chaque sous-classe lexicale différenciée. Ce prédicat peut soit résumer le contenu signifiant (catégorie descriptive), soit en définir la fonction (catégories fonctionnelles). Cette technique de catégorisation requiert de prendre le vecteur représentant chaque classe et de le comparer à un gabarit (l'ensemble des catégories prédéfinies). Ceci peut se faire de trois manières. Une première méthode est manuelle : c'est l'analyste qui, en fonction de son répertoire propre, attribue la catégorie à assigner. Une seconde est automatique : les vecteurs des sous-groupes sont comparés (par un calcul de similarité) à une définition (en extension) d'une liste catégorielle ou d'un plan de classification, c'est-à-dire un ensemble de catégories prédéfinies. Une troisième approche procède par apprentissage : l'analyste assigne des catégories sur des échantillons de textes et le système les redistribue sur les items lexicaux ayant des contextes similaires. Dans cette recherche, nous avons voulu expérimenter une technique de catégorisation thématique fondée sur l'extraction automatique des catégories à partir des documents traités. Comme l'ont souligné entre autres Louwerse et van Peer (2002 : 4), les index thématiques (catégories thématiques prédéfinies) posent plusieurs problèmes : « The index was conceived to be a practical reference, but trying to classify tales in the […] index proved problematic. » Afin de dépasser les limites de la catégorisation automatique effectuée à partir d'ensembles de catégories thématiques prédéfinies, nous exploitons certains outils statistiques permettant de faire émerger les catégories thématiques à partir des documents analysés. Cette méthode consiste à appliquer certains critères statistiques utilisés dans les domaines du repérage de l'information (pondération distribuée, tf · idf, taux d'information, entropie, etc.) (Salton, 1989 ; Yates et Ribiero, 1999) à chacune des sous-classes lexicales différenciées afin d'identifier au sein de chacune de ces sous-classes les termes les plus significatifs pouvant (suite à une évaluation de l'utilisateur) servir d'étiquette thématique pour la découverte des principaux thèmes d'un corpus.

6) Projection des catégories thématiques sur le texte. Une fois la catégorisation thématique effectuée, chaque domaine d'information peut se voir étiquetée automatiquement avec les étiquettes thématiques. Le texte est alors soumis à des analyses classiques soit qualitatives (regroupements, listes, arbres, graphes, etc.) soit quantitatives (statistiques, etc.), mais, cette fois, ce sont les catégories thématiques qui en sont l'objet.

7) La découverte, la navigation et la visualisation des thèmes identifiés. Pour assister l'analyse et l'interprétation des résultats, il est de plus en plus utile d'offrir aux analystes des moyens de visualiser de manière ergonomique ces classes et les relations entre les thèmes. Ce sont, comme le dit Barry (1998), des « *mind mapping tools* » ou des cartographies cognitives du contenu thématique des textes. Diverses technologies commencent à apparaître pour

assister ce type d'analyse (Spence, 2000 ; Fayyad, Grinstein et Wierse, 2001). Cette dimension ergonomique de représentation est un atout précieux dans le soutien de l'activité interprétative du chercheur.

5. Expérimentation

Dans le cadre de l'expérimentation présentée, nous avons appliqué la méthodologie décrite précédemment à un texte spécifique afin d'en explorer la pertinence. Nous présentons ici les résultats obtenus à partir du *Discours de la méthode* de Descartes.

Le lexique initial du *Discours de la méthode* est constitué de 2914 termes dont la fréquence varie entre 1 et 925 occurrences. Nous avons effectué un filtrage du lexique où furent supprimés les termes fonctionnels et ceux dont la fréquence a été jugée non pertinente pour la classification (c'est-à-dire les termes trop fréquents et pas assez fréquents). De plus, le lexique fut lemmatisé. Suite à ces opérations, le lexique épuré fut composé de 306 termes. La fréquence des termes retenus varie entre 5 et 56 apparitions. De plus, nous avons privilégié une segmentation par mots, à raison de 150 mots par segment. Pour la classification, nous avons utilisé le classifieur neuronal ART1.

6. Résultats

Le processus de segmentation a permis de découper le corpus initial en 154 segments de 150 mots. En utilisant le classifieur ART1, ces 154 segments furent regroupés en 83 classes. D'un point de vue strictement technique, la qualité des résultats de cette classification (moyenne de 1.86 segment par classe) est manifestement discutable. Cependant, l'objectif de notre recherche ne consiste pas à valider le processus de classification, mais bien la fécondité de notre démarche à l'égard d'un processus interprétatif beaucoup plus complexe lié à l'analyse thématique des données textuelles. Dans cette optique, les résultats obtenus s'avèrent acceptables afin d'atteindre, comme nous le verrons, notre objectif.

L'identification des principaux thèmes présents dans le corpus est effectuée en appliquant certains calculs statistiques permettant de faire émerger les catégories thématiques à partir des documents analysés. Dans le cadre de la présente expérimentation, nous avons uniquement appliqué la formule classique tf · idf (« *term frequency · inverse document frequency* ») (Salton, 1989) au lexique de chaque classe. Le principe de ce calcul peut être formulé de la manière suivante : un terme sera d'autant meilleur pour représenter le contenu d'une classe s'il est à la fois fréquent dans cette classe et rare dans l'ensemble des classes à analyser. La fréquence inverse du document, idf = log(N/n), où N est le nombre total de documents et n est le nombre de documents contenant le terme, vient donc modérer ou accentuer l'importance de la fréquence de chaque terme. Ainsi, ce calcul est utilisé, dans le cadre de notre analyse, afin d'extraire les termes les plus représentatifs des classes obtenues. Les termes retenus suite à ce calcul sont alors attribués comme « étiquette thématique » à leur classe respective.

Les processus de découverte et de navigation thématique débutent par le choix d'un terme particulier présent dans le corpus à analyser. Ce choix est effectué en fonction des intérêts – de recherche ou de lecture – du chercheur. À titre d'illustration, en sélectionnant le terme « connaissance », on constate que ce dernier opère dans plusieurs regroupements de segments. Ce terme est en effet présent dans des classes de segments (desquels la chaîne de traitement utilisée nous a permis d'extraire le lexique) où se retrouvent les autres termes suivants : « animal, artère, bête, branche, cave, chaleur, cœur, concavité, mouvement, organe, poumon, sang, veine, etc. » (classes 22, 41, 44 et 46, segments 94, 93, 96, 97, 103, 104 et 112), « air,

astre, ciel, lumière, matière, monde, terre, etc. » (classes 1, 24, 38 et 40, segments 1, 49, 80, 81, 84 et 86), « démonstration, entendement, géomètre, mathématique, méthode, philosophie, science, vrai, etc. » (classes 6 et 17, segments 32 et 37), « âme, astre, certitude, dieu, entendement, esprit, existence, idée, grand, parfaire, etc. » (classes 31, 32 et 34, segments 63, 64, 66, 67, 70, 71 et 73). Ces informations nous indiquent que le terme retenu opère dans plusieurs contextes différents. Compte tenu du vocabulaire spécifique de chacun de ces contextes (représentés par des classes de segments), nous pouvons affirmer que chacun de ces regroupements représente en fait un thème présent dans le *Discours de la méthode*. Le passage de la liste des termes du lexique de chaque classe vers l'identification et l'attribution d'un thème à une classe particulière s'effectue en identifiant les termes les plus représentatifs de chaque classe (selon les valeurs obtenues par le calcul tf · idf).

Ainsi, à partir du terme retenu, il nous est possible, dans un premier temps, d'explorer les segments du *Discours de la méthode* étiquetés des termes thématiques suivants : « artère, cœur, peau, sang». En effet, sur la base de leur lexique respectif, on constate que les classes 22, 41, 44 et 46 traitent de ces aspects particuliers de la philosophie de Descartes. Mais, il nous est aussi possible, toujours à partir de ce même terme, de découvrir un tout autre thème présent dans le corpus. En effet, le regroupement constitué des classes 1, 24, 38 et 40 est, quant à lui, caractérisé par les termes suivants : « astre, ciel, matière, physique ».

Si, en contrepartie, notre analyse se dirige vers les classes 6, 17 et 31, 34, il nous est alors possible de découvrir les extraits thématiques dont les termes représentatifs sont les suivants : « entendement, démonstration, géomètre, mathématique » (classes 6 et 17) et « certitude, dieu, entendement, existence » (classes 31 et 34).

Ces résultats semblent en majeure partie concorder avec ceux obtenus lors d'expérimentations antérieures (Forest, 2002 ; Forest et Meunier, 2000) au cours desquelles les termes thématiques de chaque classe ont été inférés subjectivement. En effet, dans le cadre de ces expérimentations, les classes présentées ci-haut s'étaient vu attribuer les catégories thématiques suivantes : « biologie » (classes 22, 41, 44 et 46), « physique » (classes 1, 24, 38 et 40), « mathématique » (classes 6 et 17) et « métaphysique » (classes 31, 32 et 34). Il est a noter cependant que la catégorisation thématique, lorsque qu'effectuée manuellement par le chercheur, semble caractérisée par un plus haut niveau de généralité. Cette observation est entre autres explicable par le fait que l'analyse n'est pas limité dans sa catégorisation uniquement aux termes présents dans le lexique du corpus.

On remarque donc que le terme « connaissance » que nous avons utilisé comme éléments clefs dans un processus de découverte thématique du *Discours de la méthode* nous guide vers un premier niveau d'analyse composé de quatre axes principaux (figure 2). Ces quatre regroupements thématiques constituent d'ailleurs des éléments classiques de la pensée de Descartes (Rodis-Lewis, 1966 et 1984).

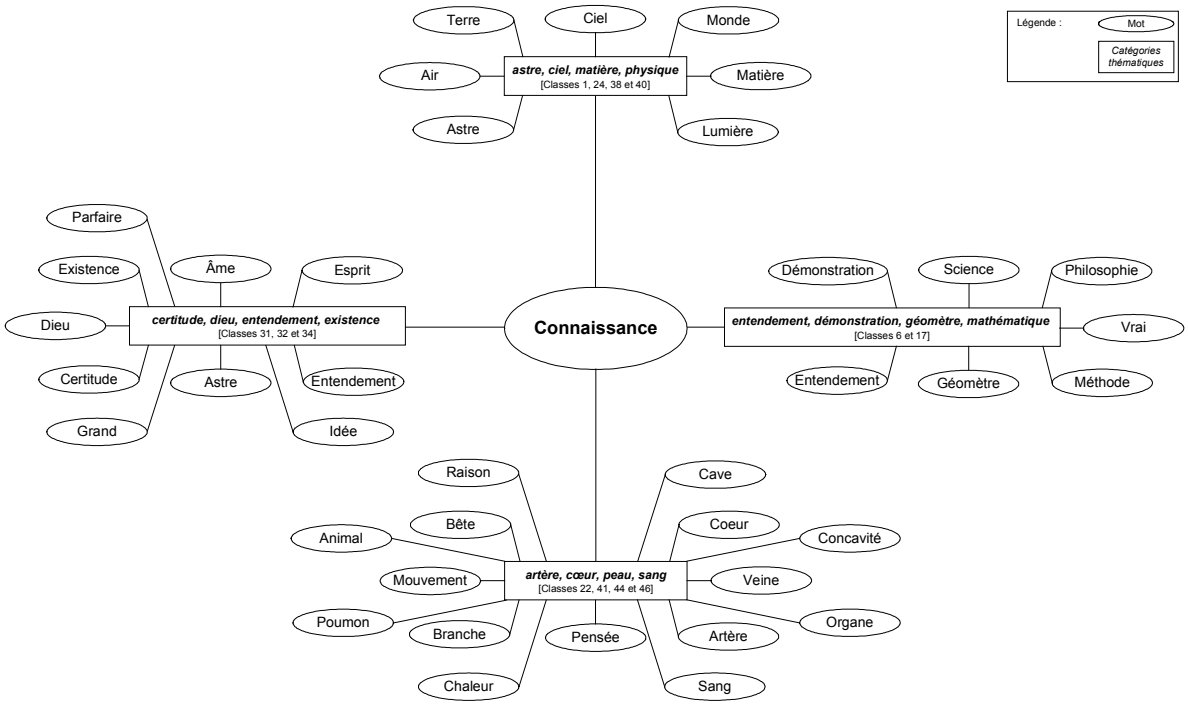


Figure 2. Échantillon des résultats thématiques obtenus à partir du terme « connaissance » du Discours de la méthode

Toutefois, certaines unités lexicales (c'est-à-dire certains mots) présentes dans une classe particulière peuvent se retrouver aussi dans une autre classe, indiquant par là qu'elles opèrent dans un autre thème. Ces mots se retrouvant dans plusieurs classes ou regroupements thématiques jouent un rôle de liaison dans l'exploration des multiples thèmes d'un corpus. C'est, en effet, sur la base de ces mots que le chercheur est amené à découvrir les contenus thématiques du corpus. Ainsi, dans une classe de départ, le lecteur peut partir d'un terme choisi et diriger son analyse vers une autre classe thématique où le même terme se retrouve, mais cette fois dans un nouveau contexte. Ce contexte, à son tour, est constitué de lexèmes qui peuvent servir de départ pour aller vers d'autres classes thématiques. C'est dans la découverte successive de plusieurs classes thématiques que consiste la navigation thématique.

Ce processus de découverte et de navigation thématique n'est pas fermé. Il se poursuit indéfiniment jusqu'à la clôture (c'est-à-dire lorsque l'analyste découvre une classe dont le lexique ne comporte pas de terme(s) servant de lien(s) vers d'autres classes thématiques) ou la saturation du parcours (c'est-à-dire lorsqu'il aura parcouru l'ensemble des thèmes de son corpus).

7. Analyse

Dans l'expérimentation présentée, la navigation thématique est perçue comme un processus de découverte et de parcours des différents thèmes présents dans un corpus textuel. Cette démarche est potentiellement multiple et fort complexe. Elle repose en dernière instance sur plusieurs choix, tant théoriques que pratiques. Mais, de manière générale, la navigation thématique consistera en un parcours caractérisé par un compromis entre, d'une part, les attentes du chercheur et, d'autre part, les indices sémiotiques présents dans le texte. Comme le souligne (Bremond, 1985 : 420) : « par quoi suis-je orienté dans la série de mes choix ? On peut

répondre : par le désir d'isoler la ou les bonnes formes du thème. Mais qu'est-ce qu'une bonne forme ? […] la bonne forme, c'est celle qui procure la satisfaction la plus grande à mon attente de lecteur […]. » Pour d'autres (Prince, 1985), cette attente du chercheur prendra le nom de « réalité extra-textuelle ». D'ailleurs plusieurs théoriciens ont noté l'importance, dans l'activité de thématisation et de découverte des contenus thématiques, de cette composante essentiellement subjective.

Dans le cadre de cette recherche, cette composante subjective se manifeste dans l'intérêt du chercheur envers certains thèmes qu'il privilégie et dans les objectifs qu'il désire atteindre lors de son analyse. Ainsi le chercheur peut, par exemple, choisir d'explorer un ou plusieurs thèmes précis du corpus, et ce dans le but d'en démontrer l'organisation, la structure, etc. Peut-être voudra-t-il explorer l'ensemble des thèmes d'un corpus afin d'orienter ou de cibler les passages qu'il voudra analyser plus en détail par la suite.

Mais d'autre part, une seconde contrainte, plus objective cette fois, entre aussi en compte dans le cadre de la tâche d'analyse et de découverte. Cette contrainte repose sur le texte à analyser. Cette composante intra-textuelle limite nécessairement la liberté de l'interprète, car elle guide inévitablement l'ensemble des analyses. En effet, malgré les intérêts et les raisons qui mènent le chercheur vers la découverte d'un thème particulier plutôt que d'un autre, le chercheur ne crée pas les thèmes dans le corpus qu'il analyse. C'est le texte qui expose, à l'aide des différents porteurs sémiotiques qu'il comporte, les thèmes sur lesquels le chercheur posera éventuellement son analyse.

8. Conclusion

L'application de techniques informatiques de catégorisation permet d'assister le chercheur dans son travail d'analyse thématique des documents (Forest, 2002 ; Forest et Meunier, 2000 ; Rossignol et Sébillot, 2002). Cependant, au-delà de la technologie et des méthodes employées, les différentes techniques de catégorisation présentent des problèmes théoriques importants. En effet, il importe de comprendre davantage la nature « catégorisante » de l'activité d'analyse thématique des documents. Il importe, dès lors, de situer le processus de catégorisation à l'égard des théories qui le sous-tendent : la logique de la classification et de la catégorisation, la sémantique cognitive, la sémantique fonctionnelle, la sémantique interprétative des textes, l'analyse de texte, etc. L'hypothèse théorique qui traverse cette recherche liée à l'utilisation des technologies pour l'analyse thématique de texte est avant tout inspirée d'une herméneutique matérielle (Rastier *et al.*, 1994), c'est-à-dire que le texte, bien que présentant une structure linguistique définie, ne révèle son contenu thématique qu'en regard d'un projet (subjectif) de lecture. L'analyse et la lecture des textes assistées par ordinateur sont des opérationnalisations informatiques des actes d'interprétation effectués sur des textes par un analyste, et dont un des moments forts est la découverte de classes de régularités sémantiques, thématiques et discursives. Comme le souligne Martin (1995 : 18), « la nature de ce que l'on pourrait appeler plus généralement l'étude thématique des textes est d'abord fonction de l'objectif visé », et Prince (1985 : 432), « thématiser un texte dépend donc non seulement du ″texte même″ mais aussi (et peut-être davantage) du thématiseur, du cadre adopté, des unités choisies, des opérations accomplies pour les harmoniser, des résumés et paraphrases effectués. » Il faut donc nécessairement lors de l'évaluation des résultats de l'application établir un compromis entre, d'une part, les techniques classiques rigoureuses d'évaluation des résultats et, d'autre part, le caractère subjectif du travail d'analyse thématique (les objectifs d'analyse thématique et les choix théoriques effectués par le chercheur).

Références

Aarnes A. et Thompson S. (1928). The types of folk-tale: a classification and bibliography. *Folklore fellows communications*, vol. (74). Suomalainen.

Alexa M. et Zuell C. (1999a). *Commonalities, difference and limitations of text analysis software: the results of a review*. ZUMA arbeitsbericht, ZUMA.

Alexa M. et Zuell C. (1999b). *A review of software for text analysis*. ZUMA arbeitsbericht, ZUMA.

Baeza-Yates R. et Ribeiro B. d. A. N. (1999). *Modern information retrieval*. ACM Press / Addison-Wesley.

Barry C.A. (1998). Choosing qualitative data analysis software: Atlas/ti and Nudist compared. *Sociological Research Online*, vol. (3/3). www.socresonline.org.uk/socresonline/3/3/4.html.

Bernard M. (1999). *Introduction aux études littéraires assistées par ordinateur*. Presses Universitaires de France.

Bremond C. (1985). Concept et thème. *Poétique*, vol. (64) : 415-423.

Bremond C., Landy J. et Pavel T. (dir. publ.) (1995). *Thematics. New approaches*. Suny Press.

Cavnar W.B. et Trenkle J.M. (1994). N-gram-based text categorization. In *Proceedings of SDAIR-94*, Las Vegas, Nevada, U.S.A., April 1994. UNLV Publications/Reprographics : 161-175.

Fayyad U., Grinstein G.G. et Wierse A. (sous la direction de) (2001). *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann Publishers.

Forest D. (2002). *Lecture et analyse de textes philosophiques assistées par ordinateur : application d'une approche classificatoire mathématique à l'analyse thématique du* Discours de la méthode *et des* Méditations métaphysiques *de Descartes*. Mémoire de maîtrise, Montréal, Université du Québec à Montréal.

Forest D. et Meunier J.-G. (2000). La classification mathématique des textes : un outil d'assistance à la lecture et à l'analyse de textes philosophiques. In *Actes des JADT 2000* : 325-329.

Giora R. (1985). Notes toward a theory of text coherence. *Poetics Today*, vol (6/4).

Grossberg S. et Carpenter G.A. (1988). The ART of adaptive pattern recognition by a self-organizing neural network. *Computer*, vol. (21/3) : 77-88.

Hockey S. (2000). *Electronic texts in the humanities*. Oxford University Press.

Jackson P. et Moulinier I. (2002). *Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization*. John Benjamins Publishing Company.

Kastberg Sjöblom M. et Brunet Ét. (2000). La thématique. Essai de repérage automatique dans l'œuvre d'un écrivain. In *Actes des JADT 2000* : 457-466.

Kintsch W et Van Dijk T.A. (1978). Toward a model of text comprehension and production. *Psychological Review*, vol. (85/5) : 363-394.

Louwerse M. et van Peer W. (sous la direction de) (2002). *Thematics: Interdisciplinary Studies*. John Benjamins Publishing Company.

Manning C.D. et Schütze H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Martin É. (1995). Thème d'étude, étude de thème. In Rastier Fr. (sous la direction de), *L'analyse thématique des données textuelles : l'exemple des sentiments*. Didier érudition : 13-24.

Meunier J.-G. (1997). La Lecture et l'Analyse de Textes Assistées par Ordinateur (LATAO) comme système de traitement d'information. *Sciences Cognitives*, vol. (22) : 211-223.

Meunier J.-G. et Torres J.M. (2000). Classphères : un réseau incrémental pour l'apprentissage non supervisé appliqué à la classification de textes. In *Actes des JADT 2000* : 365-372.

Pettersson B. (2002). Seven trends in recent thematics and a case study. In Louwerse M. et van Peer W. (sous la direction de), *Thematics: Interdisciplinary Studies*. John Benjamins Publishing Company : 237-252.

Popping R. (2000). *Computer-assisted text analysis*. Sage.

Prince G. (1985). Thématiser. *Poétique*, vol. (64) : 425-433.

Propp V. (1928/1968). *Morphology of the folktale*. Texas University Press.

Rastier Fr. (1996). *Sémantique interprétative*. Presses Universitaires de France.

Rastier, Fr. (2001). *Arts et sciences du texte*. Presses Universitaires de France.

Rastier Fr. *et al*. (sous la direction de) (1995). *L'analyse thématique des données textuelles : l'exemple des sentiments*. Didier Érudition.

Rastier Fr. *et al.* (1994). *Sémantique pour l'analyse. De la linguistique à l'informatique*. Masson.

Rimmon-Kenan S. (1985). Qu'est-ce qu'un thème ? *Poétique*, vol. (64) : 397-406.

Rodis-Lewis G. (1966). *Descartes et le rationalisme*. Presses Universitaires de France.

Rodis-Lewis G. (1984). *Descartes*. Librairie Générale Française.

Rossignol M. et Sébillot P. (2002). Automatic generation of sets of keywords for theme characterization and detection. In *Actes des JADT 2002* : 185-196.

Salton G. (1989). *Automatic Text Processing*. Addison-Wesley.

Sebastiani F. (1999). A tutorial on automated text categorisation. In Amandi A. et Zunino A. (sous la direction de), *Proceedings of ASAI-99*, Buenos Aires : 7-35.

Sebastiani F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, vol. (34/1) : 1-47.

Sebastiani F. (2003). Text Categorization. In Zanasi A. (dir. publ.), *Text Mining and its Applications*. WIT Press.

Sollors W. (1993). *The return of thematic criticism*. Harvard University Press.

Spence R. et Press A. (2000). *Information visualization*. Addison-Wesley.

Thomashevsky B. (1925). Thematics. In Lemon L.T. et Reis M.J. (Eds) (1965), *Russian formalist criticism*. University of Nebraska Press : 61-98.

Thompson S. (1946). *The folktale*. Berkeley : University of California Press.

Van Dijk T.A. (1972). *Some aspects of text grammars. A study in theoretical linguistics and poetics.* Mouton.

Van Dijk T.A. et Kintsch W. (1983). *Strategies of discourse comprehension*. Academic Press.

Some relationships between Kleinberg's hubs and authorities, correspondence analysis, and the Salsa algorithm

François Fouss[1], Jean-Michel Renders[2], Marco Saerens[1]

[1]ISYS Unit, IAG – Université catholique de Louvain – Place des Doyens 1 – 1348 Louvain-la-Neuve – Belgique
[2]Xerox Research Center Europe, Chemin de Maupertuis 6 – 38240 Meylan (Grenoble) – France

{saerens,fouss}@isys.ucl.ac.be, jean-michel.renders@xrce.xerox.com

Abstract

In this work, we show that Kleinberg's hubs and authorities model is closely related to both correspondence analysis, a well-known multivariate statistical technique, and a particular Markov chain model of navigation through the web. The only difference between correspondence analysis and Kleinberg's method is the use of the **average** value of the hubs (authorities) scores for computing the authorities (hubs) scores, instead of the **sum** for Kleinberg's method. We also show that correspondence analysis and our Markov model are related to SALSA, a variant of Kleinberg's model. We finally suggest that the Markov model could easily be extended to the analysis of more general structures, such as relational databases.

1. Introduction

Exploiting the graph structure of large document repositories, such as the web environment, is one of the main challenges of computer science and data mining today. In this respect, Kleinberg's proposition to distinguish web pages that are hubs and authorities (see Kleinberg, 1999); called the HITS algorithm) has been well-received in the community.

In this paper, we show that Kleinberg's hubs and authorities procedure (Kleinberg, 1999) is closely related to both correspondence analysis (see for instance Greenacre, 1984 ; Lebart *et al.*, 1995), a well-known multivariate statistical analysis technique, and a particular Markov chain model of navigation through the web. We further show that correspondence analysis and the Markov model are related to SALSA (Lempel and Moran, 2001), a variant of Kleinberg's model. This puts new lights on the interpretation of Kleinberg's procedure since correspondence analysis has a number of interesting properties that makes it well suited for the analysis of frequency tables. On the other hand, the Markov model can easily be extended to more general structures, such as relational databases

In section 2, we briefly introduce the basics of Kleinberg's procedure for ranking query's results. In section 3, we introduce correspondence analysis and relate it to Kleinberg's procedure while, in section 4, we introduce a Markov model of web navigation and relate it to both correspondence analysis and Kleinberg's procedure.

2. Kleinberg's procedure

In 1999, Kleinberg introduced a procedure for identifying web pages that are good hubs or good authorities, in response to a given query. The following example is often mentioned.

When considering the query "automobile makers", the home pages of Ford, Toyota and other car makers are considered as good authorities, while web pages that list these home pages are good hubs.

2.1. *The updating rule*

To identify good hubs and authorities, Kleinberg's procedure exploits the graph structure of the web. Each web page is a node and a link from page a to page b is represented by a directed edge from node a to node b. When introducing a query, the procedure first constructs a focused subgraph G, and then computes hubs and authorities scores for each node of G. Let n be the number of nodes of G. We now briefly describe how these scores are computed. Let $\mathbf{W}$ be the adjacency matrix of the subgraph G; that is, element w_{ij} (row i, column j) of matrix $\mathbf{W}$ is equal to 1 if and only if node (web page) i contains a link to node (web page) j; otherwise, $w_{ij} = 0$. We respectively denote by $\mathbf{x}^h$ and $\mathbf{x}^a$ the hubs and authorities $n \times 1$ column vector scores corresponding to each node of the subgraph.

Kleinberg uses an iterative updating rule in order to compute the scores. Initial scores at $k = 0$ are all set to 1, i.e. $\mathbf{x}^h = \mathbf{x}^a = \mathbf{1}$ where $\mathbf{1} = [1, 1, \ldots, 1]^{\mathrm{T}}$ is a $n \times 1$ column vector made of 1. Then, the following mutually reinforcing rule is used: The Hub score for node i, x_i^h, is set equal to the normalized sum of the authority scores of all nodes pointed by i and, similarly, the authority score of node j, x_j^a, is set equal to the normalized sum of hub scores of all nodes pointing to j. This corresponds to the following updating rule:

$$\mathbf{x}^h(k+1) = \frac{\mathbf{W}\mathbf{x}^a(k)}{\|\mathbf{W}\mathbf{x}^a(k)\|_2} \quad (1)$$

$$\mathbf{x}^a(k+1) = \frac{\mathbf{W}^{\mathrm{T}}\mathbf{x}^h(k)}{\|\mathbf{W}^{\mathrm{T}}\mathbf{x}^h(k)\|_2} \quad (2)$$

where $\|\mathbf{x}\|_2$ is the Euclidian norm, $\|\mathbf{x}\|_2 = (\mathbf{x}^{\mathrm{T}}\mathbf{x})^{1/2}$.

2.2. *An eigenvalue/eigenvector problem*

Kleinberg (1999) showed that when following this update rule, $\mathbf{x}^h$ converges to the normalized principal (or dominant) right eigenvector of the symmetric matrix $\mathbf{W}\mathbf{W}^{\mathrm{T}}$, while $\mathbf{x}^a$ converges to the normalized principal eigenvector of the symmetric matrix $\mathbf{W}^{\mathrm{T}}\mathbf{W}$, provided that the eigenvalues are distinct.

Indeed, the equations (1), (2) result from the application of the power method, an iterative numerical method for computing the dominant eigenvector of a symmetric matrix (Golub and Loan, 1996), to the following eigenvalue problem:

$$\mathbf{x}^h \propto \mathbf{W}\mathbf{x}^a \Rightarrow \mathbf{x}^h = \mu\mathbf{W}\mathbf{x}^a \quad (3)$$

$$\mathbf{x}^a \propto \mathbf{W}^{\mathrm{T}}\mathbf{x}^h \Rightarrow \mathbf{x}^a = \eta\mathbf{W}^{\mathrm{T}}\mathbf{x}^h \quad (4)$$

where $\propto$ means "proportional to" and T is the matrix transpose. Or, equivalently,

$$x_i^h \propto \sum_{j=1}^{n} w_{ij}x_j^a \Rightarrow x_i^h = \mu\sum_{j=1}^{n} w_{ij}x_j^a \quad (5)$$

$$x_j^a \propto \sum_{i=1}^{n} w_{ij}x_i^h \Rightarrow x_j^a = \eta\sum_{i=1}^{n} w_{ij}x_i^h \quad (6)$$

Meaning that each hub node, i, is given a score, x_i^h, that is proportional to the sum of the authorities nodes scores to which it links to. Symmetrically, to each authorities node, j, we allocate a score, x_j^a, which is proportional to the sum of the hubs nodes scores that point to it. By substituting (3) in (4) and vice-versa, we easily obtain

$$\begin{aligned} \mathbf{x}^h &= \mu\eta\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{x}^h = \lambda\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{x}^h \\ \mathbf{x}^a &= \mu\eta\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{x}^a = \lambda\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{x}^a \end{aligned}$$

which is an eigenvalue/eigenvector problem.

2.3. *A variant of Kleinberg's procedure*

Many extensions of the updating rules (1), (2) were proposed. For instance, in Lempel and Moran (2001) (the SALSA algorithm), the authors propose to normalise the matrices $\mathbf{W}$ and $\mathbf{W}^{\mathrm{T}}$ in (3) and (4) so that the new matrices verify $\mathbf{W}'\mathbf{1} = \mathbf{1}$ and $\left(\mathbf{W}^{\mathrm{T}}\right)'\mathbf{1} = \mathbf{1}$ (the sum of the elements of each row of $\mathbf{W}'$ and $\left(\mathbf{W}^{\mathrm{T}}\right)'$ is 1). In this case, (3) and (4) can be rewritten as

$$x_i^h \propto \sum_{j=1}^{n} w'_{ij}x_j^a = \frac{\sum_{j=1}^{n} w_{ij}x_j^a}{w_{i.}} \text{, where } w_{i.} = \sum_{j=1}^{n} w_{ij} \tag{7}$$

$$x_j^a \propto \sum_{i=1}^{n} w'_{ij}x_i^h = \frac{\sum_{i=1}^{n} w_{ij}x_i^h}{w_{.j}} \text{, where } w_{.j} = \sum_{i=1}^{n} w_{ij} \tag{8}$$

This normalization has the effect that nodes (web pages) having a large number of links are not privileged with respect to nodes having a small number of links. In the next section,we will show that this variant of Kleinberg's procedure is equivalent to correspondence analysis.

3. Correspondence analysis and Kleinberg's procedure

Correspondence analysis is a standard multivariate statistical analysis technique aiming to analyse frequency tables (Greenacre, 1984; Mardia *et al.*, 1979; Lebart *et al.*, 1995).

3.1. *Correspondence analysis*

Imagine that we have a table of frequencies, $\mathbf{W}$, for which each cell, w_{ij}, represents the number of cases having both values i for the row variable and j for the column variable (we simply use the term "value" for the discrete value taken by a categorical variable). In our case, the records are the directed edges; the row variable represents the index of the origin node of the edge (hubs) and the column variable the index of the end node of the edge (authorities).

Correspondence analysis associates a score to the values of each of these variables. These scores

relate the two variables by what is called a "**reciprocal averaging**" relation (Greenacre, 1984):

$$x_i^h \propto \frac{\sum_{j=1}^{n} w_{ij} x_j^a}{w_{i.}}, \text{ where } w_{i.} = \sum_{j=1}^{n} w_{ij} \tag{9}$$

$$x_j^a \propto \frac{\sum_{i=1}^{n} w_{ij} x_i^h}{w_{.j}}, \text{ where } w_{.j} = \sum_{i=1}^{n} w_{ij} \tag{10}$$

which is exactly the same as (7) and (8). This means that each hub node, i, is given a score, x_i^h, that is proportional to the average of the authorities nodes scores to which it links to. Symmetrically, to each authorities node, j, we allocate a score, x_j^a, which is proportional to the average of the hubs nodes scores that point to it.

3.2. *Links with Kleinberg's procedure*

Notice that (9) and (10) differ from (5) and (6) only by the fact that we use the **average value** in order to compute the scores, instead of the sum.

Now, by defining the diagonal matrix $\mathbf{D}^h = \mathrm{diag}(1/w_{i.})$ and $\mathbf{D}^a = \mathrm{diag}(1/w_{.j})$ containing the number of links, we can rewrite (9) and (10) in matrix form

$$\mathbf{x}^h \propto \mathbf{D}^h \mathbf{W} \mathbf{x}^a = \mu \mathbf{D}^h \mathbf{W} \mathbf{x}^a \tag{11}$$

$$\mathbf{x}^a \propto \mathbf{D}^a \mathbf{W}^{\mathrm{T}} \mathbf{x}^h = \eta \mathbf{D}^a \mathbf{W}^{\mathrm{T}} \mathbf{x}^h \tag{12}$$

In the language of correspondence analysis, the row vectors of $\mathbf{D}^h\mathbf{W}$ are the hub **profiles**, while the row vectors of $\mathbf{D}^a\mathbf{W}^{\mathrm{T}}$ are the authorities **profiles**. These vectors sum to one.

Now, from (11), (12), we easily find

$$\mathbf{x}^h = \mu\eta \mathbf{D}^h \mathbf{W} \mathbf{D}^a \mathbf{W}^{\mathrm{T}} \mathbf{x}^h = \lambda \mathbf{D}^h \mathbf{W} \mathbf{D}^a \mathbf{W}^{\mathrm{T}} \mathbf{x}^h \tag{13}$$

$$\mathbf{x}^a = \mu\eta \mathbf{D}^a \mathbf{W}^{\mathrm{T}} \mathbf{D}^h \mathbf{W} \mathbf{x}^a = \lambda \mathbf{D}^a \mathbf{W}^{\mathrm{T}} \mathbf{D}^h \mathbf{W} \mathbf{x}^a \tag{14}$$

Correspondence analysis computes the subdominant right eigenvector of $\mathbf{D}^h\mathbf{W}\mathbf{D}^a\mathbf{W}^{\mathrm{T}}$ and $\mathbf{D}^a\mathbf{W}^{\mathrm{T}}\mathbf{D}^h\mathbf{W}$. Indeed, it can be shown that the right principal eigenvector (the largest one) is a trivial one, $[1, 1, \ldots, 1]^{\mathrm{T}}$ with eigenvalue $\lambda = 1$ (all the other eigenvalues are positive and smaller that 1; see Greenacre, 1984) since the column values of $\mathbf{D}^h\mathbf{W}\mathbf{D}^a\mathbf{W}^{\mathrm{T}}$ (respectively $\mathbf{D}^a\mathbf{W}^{\mathrm{T}}\mathbf{D}^h\mathbf{W}$) sum to one for each row. Therefore, correspondence analysis computes the second largest eigenvalue, called the **subdominant eigenvalue**, as well as the corresponding eigenvector.

In standard correspondence analysis, this subdominant right eigenvector has several interesting interpretations in terms of "optimal scaling", of the "best approximation" to the original matrix or of the linear combinations of the two sets of values that are "maximally correlated", etc. (see for instance Greenacre, 1984; Lebart *et al.*, 1995). With respect to this last interpretation, it can be shown that correspondance analysis computes $\max_{\mathbf{a},\mathbf{b}} \left[\mathrm{corr}\left(\mathbf{W}\mathbf{a}, \mathbf{W}^{\mathrm{T}}\mathbf{b}\right)\right]$.

The next eigenvectors can be computed as well; they are related to the proportion of chi-square computed on the original table of frequencies that can be explained by the m first eigenvectors. Correspondence analysis is therefore often considered as an "equivalent" of principal components analysis for frequency tables.

4. A Markov chain model of web navigation

We now introduce a Markov chain model of random web navigation that is equivalent to correspondence analysis, and therefore closely related to Kleinberg's procedure. It therefore provides a new interpretation for both correspondence analysis and Kleinberg's procedure. Notice that this random walk model is very similar to the one proposed in Lempel and Moran (2001) (for other random walk models, see also the PageRank system (Page *et al.*, 1998; Ng *et al.*, 2001), and has some interesting links with diffusion kernels (Kondor and Lafferty, 2002), or with a measure of similarity between graphs vertices that was proposed in Blondel and Senellart (2002), which we are currently investigating.

4.1. *Definition of the Markov chain*

We first define a Markov chain in the following way. We associate a state of the Markov chain to every hub and every authority node ($2n$ in total); we also define a random variable, $s(k)$, representing the state of the Markov model at time step k . Moreover, let s^h be the subset of states that are hubs and s^a be the subset of states that are authorities. We say that $s^h(k) = i$ (respectively $s^a(k) = i$) when the Markov chain is in the state corresponding to the i^{th} hub (authority) at time step k. As in Lempel and Moran (2001), we define a random walk on these states by the following single-step transition probabilities

$$\mathrm{P}(s^h(k+1) = i | s^a(k) = j) = \frac{w_{ij}}{w_{.j}} \text{ , where } w_{.j} = \sum_{i=1}^{n} w_{ij} \tag{15}$$

$$\mathrm{P}(s^a(k+1) = j | s^h(k) = i) = \frac{w_{ij}}{w_{i.}} \text{ , where } w_{i.} = \sum_{j=1}^{n} w_{ij} \tag{16}$$

$$\mathrm{P}(s^a(k+1) = j | s^a(k) = i) = \mathrm{P}(s^h(k+1) = j | s^h(k) = i) = 0 \text{ , for all } i, j \tag{17}$$

In other words, to any hub page, $s^h(k) = i$, we associate a non-zero probability of jumping to an authority page, $s^a(k+1) = j$, pointed by the hub page (equation 16), which is inversely proportional to the number of directed edges leaving $s^h(k) = i$. Symmetrically, to any authority page $s^a(k) = i$, we associate a non-zero probability of jumping to a hub page $s^h(k+1) = j$ pointing to the authority page (equation 15), which is inversely proportional to the number of directed edges pointing to $s^a(k) = i$. We suppose that the Markov chain is irreducible, that is, every state can be reached from any other state. If this is not the case, the Markov chain can be decomposed into closed sets of states which are completely independent (there is no communication between them), each closed set being irreducible. In this situation, our analysis can be performed on these closed sets instead of the full Markov chain.

Now, if we denote the probability of being in a state by $x_i^h(k) = \mathrm{P}(s^h(k) = i)$ and $x_i^a(k) = \mathrm{P}(s^a(k) = i)$, and we define $\mathbf{P}^h$ as the transition matrix whose elements are $p_{ij}^h = \mathrm{P}(s^h(k+1) = j | s^a(k) = i)$ and $\mathbf{P}^a$ as the transition matrix whose elements are $p_{ij}^a = \mathrm{P}(s^a(k+1) = j | s^h(k) = i)$, from equations (15) and (16),

$$\begin{aligned} \mathbf{P}^h &= \mathbf{D}^a \mathbf{W}^{\mathrm{T}} \\ \mathbf{P}^a &= \mathbf{D}^h \mathbf{W} \end{aligned}$$

The Markov model is characterized by

$$\begin{aligned} x_i^h(k+1) &= \mathrm{P}(s^h(k+1)=i) = \sum_{j=1}^{n} \mathrm{P}(s^h(k+1)=i|s^a(k)=j)\, x_j^a(k) \\ &= \sum_{j=1}^{n} p_{ji}^h\, x_j^a(k) \\ x_i^a(k+1) &= \mathrm{P}(s^a(k+1)=i) = \sum_{j=1}^{n} \mathrm{P}(s^a(k+1)=i|s^h(k)=j)\, x_j^h(k) \\ &= \sum_{j=1}^{n} p_{ji}^a\, x_j^h(k) \end{aligned}$$

Or, in matrix form,

$$\mathbf{x}^h(k+1) = (\mathbf{P}^h)^{\mathrm{T}}\mathbf{x}^a(k) \quad (18)$$
$$\mathbf{x}^a(k+1) = (\mathbf{P}^a)^{\mathrm{T}}\mathbf{x}^h(k) \quad (19)$$

It is easy to observe that the Markov chain is periodical with period 2: each hub (authority) state could potentially be reached in one jump from an authority (hub) state but certainly not from any other hub (authority) state. In this case, the set of hubs (authorities) corresponds to a subset which itself is an irreducible and aperiodic Markov chain whose evolution is given by

$$\mathbf{x}^h(k+2) = (\mathbf{P}^h)^{\mathrm{T}}(\mathbf{P}^a)^{\mathrm{T}}\mathbf{x}^h(k) = (\mathbf{Q}^h)^{\mathrm{T}}\mathbf{x}^h(k) \quad (20)$$
$$\mathbf{x}^a(k+2) = (\mathbf{P}^a)^{\mathrm{T}}(\mathbf{P}^h)^{\mathrm{T}}\mathbf{x}^a(k) = (\mathbf{Q}^a)^{\mathrm{T}}\mathbf{x}^a(k) \quad (21)$$

where $\mathbf{Q}^h$ and $\mathbf{Q}^a$ are the transition matrices of the corresponding Markov models for the hubs and authorities. This Markov chain is aperiodic since each link (corresponding to a transition) can be followed in both directions (from hub to authority and from authority to hub) so that, when starting from a state, we can always return to this state in two steps. Hence, all the diagonal elements of $\mathbf{Q}^h$ and $\mathbf{Q}^a$ are non-zero and the Markov chain is aperiodic.

Therefore, the transition matrices of the corresponding Markov chains are

$$\mathbf{Q}^h = \mathbf{P}^a\mathbf{P}^h = \mathbf{D}^h\mathbf{W}\mathbf{D}^a\mathbf{W}^{\mathrm{T}} \quad (22)$$
$$\mathbf{Q}^a = \mathbf{P}^h\mathbf{P}^a = \mathbf{D}^a\mathbf{W}^{\mathrm{T}}\mathbf{D}^h\mathbf{W} \quad (23)$$

The matrices appearing in these equations are equivalent to the ones appearing in (13), (14). We will now see that the subdominant right eigenvectors of these matrices (which are computed in correspondence analysis) have an interesting interpretation in terms of the distance to the "steady state" of the Markov chain.

4.2. *Interpretation of the subdominant right eigenvector of* $\mathbf{Q}^h$, $\mathbf{Q}^a$

Now, it is well-know that the subdominant right eigenvector of the transition matrix, $\mathbf{Q}$, of an irreducible, aperiodic, Markov chain measures the departure of each state from the "equilibrium position" or "steady-state" probability vector (see for instance Stewart, 1994; a proof is provided in appendix 6), π, which is given by the first left eigenvector of the transition matrix $\mathbf{Q}$:

$$\mathbf{Q}^{\mathrm{T}}\pi = \pi \text{ subject to } \sum_{i=1}^{n} \pi_i = 1 \tag{24}$$

with eigenvalue $\lambda = 1$. This principal left eigenvector, π, is unique and positive and is called the "steady state" vector. This "steady state" vector, π, is the probability of finding the Markov chain in state $s = i$ in the long-run behavior:

$$\lim_{k\to\infty} \mathrm{P}(s(k) = i) = \pi_i \tag{25}$$

and is independent of the initial distribution of states at $k = 0$. It represents the probability of being in a particular state in the long-run behaviour.

In appendix 6, we show that the elements of the subdominant right eigenvector of $\mathbf{Q} = \mathbf{Q}^h$ or $\mathbf{Q}^a$ can be interpreted as a "distance" from each state to its "steady-state" value, provided by π. The states of the Markov chain are often classified or clustered by means of the values of this subdominant eigenvector as well as the few next eigenvectors.

Notice that we compute the subdominant eigenvector because the right dominant eigenvector is $\mathbf{1}$, since the sums of the columns of $\mathbf{Q}$ is one ($\mathbf{Q1} = \mathbf{1}$), so that $\mathbf{1}$ is in fact the right dominant eigenvector of $\mathbf{Q}$ with eigenvalue 1. Indeed, from the theory of nonnegative matrices, this eigenvalue is simple and is strictly larger in magnitude than the remaining eigenvalues.

4.3. *Hubs and authorities scores*

Lempel and Moran (2001), in the SALSA algorithm, propose, as hubs and authorities scores, to compute the steady-state vectors, π^h and π^a, corresponding to the hubs matrix, $\mathbf{Q}^h$, and the authorities matrix, $\mathbf{Q}^a$. We propose instead or, more precisely, in addition, to use the subdominant right eigenvector, which produces the same results as correspondence analysis, and which is often used in order to characterise the states of the Markov chain.

Indeed, in the appendix, we show that the behavior of the Markov model can be expanded into a series of left/right eigenvectors (see equation (32)). The first term is related to the steady-state vector, while the second to the time needed to reach this steady state. Eventually, the next eigenvector/eigenvalue could be computed as well; it corresponds to second-order corrections. Of course, only experimental results could confirm this choice; we are currently performing experiments investigating this issue.

4.4. *Computation of the steady-state vector*

In Lempel and Moran (2001), it is shown that, in our case, the steady-state vectors π^h and π^a are given by

$$\pi_i^h = \frac{\sum_{j=1}^{n} w_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} = \frac{w_{i.}}{w_{..}} \tag{26}$$

$$\pi_j^a = \frac{\sum_{i=1}^{n} w_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} = \frac{w_{.j}}{w_{..}} \tag{27}$$

which are the marginal proportions of the frequency table $\mathbf{W}$. It is therefore not necessary to compute the eigenvector. This is in fact a standard result for a random walk on a graph (see for instance Ross, 1996). From the theory of finite-state irreducible and aperiodic Markov chains (Cinlar, 1975; Bremaud, 1999), we know that there exists exactly one limiting density obeying (26, 27), given by (24).

We saw that the subdominant eigenvector of $\mathbf{Q}^h$ and $\mathbf{Q}^a$ represents the departure from the marginal proportions of the frequency table. This is quite similar to correspondence analysis where the same eigenvectors are interpreted in terms of a chi-square distance to the "independence model", for which we assume $\mathrm{P}(s^a(k) = i, s^h(k) = j) = (w_{i.}w_{.j})/(w_{..}^2)$ (Greenacre, 1984; Lebart *et al.*, 1995).

Moreover, the markov chain is reversible (see Ross, 1996). This has important implications. For instance, it is known that all the eigenvalues of the transition matrix of a reversible Markov chain are real and distinct (see for instance Bremaud, 1999).

4.5. *The full Markov chain model*

Finally, notice that (18), (19) can be rewritten as

$$\begin{bmatrix} \mathbf{x}^h(k+1) \\ \mathbf{x}^a(k+1) \end{bmatrix} = \begin{bmatrix} 0 & (\mathbf{P}^h)^{\mathrm{T}} \\ (\mathbf{P}^a)^{\mathrm{T}} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}^h(k) \\ \mathbf{x}^a(k) \end{bmatrix} \tag{28}$$

so that the transition matrix of the full Markov chain with state vector $\left[(\mathbf{x}^h)^{\mathrm{T}}, (\mathbf{x}^a)^{\mathrm{T}}\right]^{\mathrm{T}}$ is

$$\mathbf{P} = \begin{bmatrix} 0 & \mathbf{P}^a \\ \mathbf{P}^h & 0 \end{bmatrix}$$

If $\mathbf{u}^h$ and $\mathbf{u}^a$ are respectively eigenvectors of $\mathbf{Q}^h$ and $\mathbf{Q}^a$, it is easy to show that the transition matrix $\mathbf{P}^2$ has eigenvectors $\left[(\mathbf{u}^h)^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}}\right]^{\mathrm{T}}$ and $\left[\mathbf{0}^{\mathrm{T}}, (\mathbf{u}^a)^{\mathrm{T}}\right]^{\mathrm{T}}$.

5. Possible extensions of Kleinberg's model, based on correspondence analysis and the Markov chain model

The relationships between Kleinberg's procedure, correpondence analysis and Markov models suggest extensions of Kleinberg's HITS algorithm in three differents directions:

1. The proposed random walk procedure can easily be extended to the analysis of more complex structures, such as relational databases. Here is a sketch of the main idea. To each record of a table, we associate a state; each table corresponding to a subset of states. For each relation between two tables, we associate a set of transitions. This way, we can define a random walk model and compute interesting information from it, such as the distance to the steady state, the average first passage time from one state to another, etc, with the objective of computing similarities between states. This would allow us to compute similarities between records (states) of a same table as well as records (states) belonging to different tables of the database. These developments, as well as an experimental validation of the method, will be the subject of a forthcoming paper. In the same vein, extensions of Kleinberg's procedure were also proposed in Blondel and VanDooren (2002), with the objective of computing similarities between nodes of a graph. We also intend to generalize correspondence analysis to multiple correspondence analysis (correspondence analysis applied to multiple tables) by using the concept of stochastic complementation (Meyer, 1989).

2. Instead of computing only the first eigenvalue/eigenvector, we could additionally compute the next eigenvevalues/eigenvectors (as done in correspondence analysis) that could also provide additional information, as already suggested by Kleinberg himself. With this respect, notice that in correspondence analysis, each eigenvalue is directly related to the amount of chi-square accounted for the corresponding eigenvector. This way, we extract as many eigenvectors as there are significant eigenvalues.

3. The various interpretations of correspondence analysis put new light to Kleinberg's procedure. Moreover, extensions of correspondence analysis are available (for instance multiple correspondence analysis); these extensions could eventually be applied in order to analyse the graph structure of the web.

6. Conclusions and further work

We showed that Kleinberg's method for computing hubs and authorities scores is closely related to correspondence analysis, a well-known multivariate statistical analysis method. This allows us to give some new interpretations to Kleinberg's method. Also, this suggests to extract the next eigenvalues/eigenvectors as well, since these could provide some additional information (as is done in correspondence analysis).

We then introduce a Markov random walk model of navigation through the web, and we show that this model is also equivalent to correspondence analysis. This random walk procedure has an important advantage: it can easily be extended to more complex structures, such as relational databases. These developments will be the subject of a forthcoming paper. In the same vein, extensions of Kleinberg's procedure were also proposed in Blondel and VanDooren (2002), with the objective of computing similarities between nodes of a graph.

Acknowledgments

We thank Prof. Vincent Blondel, from the "Département d'Ingéniérie Mathematique" of the Université catholique de Louvain, Prof. Guy Latouche, Prof. Guy Louchart and Dr. Pascal Franck, both from the Université Libre de Bruxelles, for insightful discussions.

APPENDIX: PROOF OF THE MAIN RESULTS

Appendix: Distance to the steady state vector

In this appendix, we show that the entries of the subdominant right eigenvector of the transition matrix $\mathbf{Q}$ of a finite, aperiodic, irreducible, reversible, Markov chain can be interpreted as a distance to the "steady-state" probability vector, π. From (22), (23), we can easily show that $\mathbf{Q}$ has a positive real spectrum so that all its eigenvalues are positive real and its eigenvectors are real. Moreover, since $\mathbf{Q}$ is stochastic nonnegative, all the eigenvalues are ≤ 1, and the eigenvalue 1 has multiplicity one. The proof is adapted from Papoulis and Pillai, 2002; Stewart, 1994; Bremaud, 1999.

Let $\mathbf{e}_i = [0, 0, \ldots, 1, \ldots, 0, 0]^{\mathrm{T}}$ be the column vector with i^{th} component equal to 1, all others being equal to 0. $\mathbf{e}_i$ will denote that, initially, the system starts in state i. Since $\mathbf{Q}$ is aperiodic, irreducible, and reversible, we know that $\mathbf{Q}$ has n simple, real, distinct, nonzero eigenvalues.

After one time step, the probability density of finding the system in one state is (see (18),(19))

$$\mathbf{x}(1) = \mathbf{Q}^{\mathrm{T}}\mathbf{e}_i$$

After k steps, we have

$$\mathbf{x}(k) = \left(\mathbf{Q}^{\mathrm{T}}\right)^k \mathbf{e}_i$$

The idea is to compute the distance

$$d_i(k) = \left\| \left(\mathbf{Q}^{\mathrm{T}}\right)^k \mathbf{e}_i - \pi \right\|_2 \tag{29}$$

in order to have an idea of the rate of convergence to the steady state when starting from a particular state $s = i$.

Let $(\lambda_i, \mathbf{u}_i)$, $i = 1, 2, \ldots n$ represent the n right eigenvalue-eigenvectors pairs of $\mathbf{Q}$ in decreasing order of importance (of modulus). Thus

$$\mathbf{Q}\mathbf{U} = \mathbf{U}\mathbf{\Lambda} \tag{30}$$

where $\mathbf{U}$ is the $n \times n$ matrix made of the column vectors $\mathbf{u}_i$ which form a basis of $\Re^n$: $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n]$, and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_i)$.

From (30),

$$\mathbf{Q} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1} = \mathbf{U}\mathbf{\Lambda}\mathbf{V} \tag{31}$$

where we set $\mathbf{V} = \mathbf{U}^{-1}$. We therefore obtain $\mathbf{V}\mathbf{Q} = \mathbf{\Lambda}\mathbf{V}$, where $\mathbf{V} = \mathbf{U}^{-1} = \begin{bmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_n^{\mathrm{T}} \end{bmatrix}$, so that the column vectors $\mathbf{v}_i$ are the left eigenvectors of $\mathbf{Q}$, $\mathbf{v}_i^{\mathrm{T}}\mathbf{Q} = \lambda_i \mathbf{v}_i^{\mathrm{T}}$. Moreover, since $\mathbf{V}\mathbf{U} = \mathbf{I}$, we have $\mathbf{v}_i^{\mathrm{T}}\mathbf{u}_j = \delta_{ij}$.

Hence from (31),

$$\begin{aligned}
\mathbf{Q}^k &= \mathbf{U}\mathbf{\Lambda}^k\mathbf{V} \\
&= \sum_{i=1}^{n} \lambda_i^k \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}} \\
&= \mathbf{1}\pi^{\mathrm{T}} + \sum_{i=2}^{n} \lambda_i^k \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}} \\
&= \mathbf{1}\pi^{\mathrm{T}} + \lambda_2^k \mathbf{u}_2 \mathbf{v}_2^{\mathrm{T}} + O\left((n-2)\lambda_3^k\right)
\end{aligned} \tag{32}$$

since $\lambda_i < 1$ for $i > 1$ and the eigenvalues/eigenvectors are ordered in decreasing order of eigenvalue magnitude. This development of the transition matrix is often called the spectral decomposition of $\mathbf{Q}^k$.

Let us now return to (29)

$$\begin{aligned}
d_i(k) &= \left\| \left(\mathbf{Q}^{\mathrm{T}}\right)^k \mathbf{e}_i - \pi \right\|_2 \\
&\simeq \left\| \left(\pi\mathbf{1}^{\mathrm{T}} + \lambda_2^k \mathbf{v}_2 \mathbf{u}_2^{\mathrm{T}}\right) \mathbf{e}_i - \pi \right\|_2 \\
&\simeq \left\| \pi + \lambda_2^k \mathbf{v}_2 \mathbf{u}_2{}^{\mathrm{T}} \mathbf{e}_i - \pi \right\|_2 \\
&\simeq \lambda_2^k \left\| \mathbf{v}_2 \right\|_2 u_{2i}
\end{aligned}$$

where u_{2i} is i^{th} component of $\mathbf{u}_2$. Since the only term that depends on the initial state, i, is u_{2i}, the eigenvector $\mathbf{u}_2$ can be interpreted as a distance to the steady-state vector.

References

Blondel V.D. and Dooren P.V. (2002). A measure of similarity between graph vertices, with application to synonym extraction and web searching. *Technical Report UCL 02-50*. Université catholique de Louvain.

Blondel V.D. and Senellart P.P. (2002). Automatic extraction of synonyms in a dictionary. In *Proceedings of the SIAM Text Mining Workshop*, Airlington, Virginia.

Bremaud P. (1999). *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer-Verlag.

Cinlar E. (1975). *Introduction to Stochastic Processes*. Prentice-Hall.

Golub G.H. and Loan C.F.V. (1996). *Matrix Computations* (3th Ed.). The Johns Hopkins University Press.

Greenacre M.J. (1984) *Theory and Applications of Correspondence Analysis*. Academic Press.

Kleinberg J.M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, vol. (46/5): 604-632.

Kondor R.I. and Lafferty J. (2002). Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the 19th International Conference on Machine Learning*.

Lebart L., Morineau A. and P.M. (1995). *Statistique Exploratoire Multidimensionnelle*. Dunod.

Lempel R. and Moran S. (2001). Salsa: The stochastic approach for link-structure analysis. In *ACM Transactions on Information Systems*, vol. (19/2): 131-160.

Mardia K., Kent J. and Bibby J. (1979). *Multivariate Analysis*. Academic Press.

Meyer C. (1989). Stochastic complementation, uncoupling markov chains, and the theory of nearly reducible systems. *SIAM Review*, vol. (31): 240-272.

Ng A.Y., Zheng A.X. and Jordan M.I. (2001). Link analysis, eigenvectors and stability. In *International Joint Conference on Artificial Intelligence (IJCAI-01)*.

Page L., Brin S., Motwani R. and Winograd T. (1998). The pagerank citation ranking: Bringing order to the web. *Technical Report, Computer System Laboratory*. Stanford University.

Papoulis A. and Pillai S.U. (2002). *Probability, Random Variables and Stochastic Processes*. McGraw-Hill.

Ross S. (1996). *Stochastic Processes* (2nd Ed.). Wiley.

Stewart W.J. (1994). *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press.

De vs *des* devant les noms précédés d'épithète en français : le problème de *petit*

Itsuko Fujimura[1], Mitsumi Uchida[2], Hiroshi Nakao[3]

[1] Université de Nagoya – 4648601 Nagoya – Japon – fujimura@cc.nagoya-u.ac.jp

[2] Université féminine d'Osaka – 5900035 Sakai – Japon

[3] Université de Aïchi – 4418522 Toyohashi – Japon

Abstract

This study discusses the alternation of the French articles *de* and *des* in noun phrases with a preposed adjective, Traditionally, it has been argued that the choice is conditioned by register and by the strength of the collocation between the adjective and the noun. In our earlier study, however, it was shown that these two factors fall short of a conclusive description of the observed facts. Notably, the particular tendencies displayed by *petit(e)s* could not be accounted for. For a comprehensive characterization of the alternation, we carried out surveys using very large corpora. Based on statistical analyses, we will demonstrate that *petit* apparently has no special preference for composing compound words and then point out two other relevant factors: (1) lexical properties of the adjective; (2) semantico--functional status of the adjectival in the discourse. These two factors together constitute "weight" of the adjectival and affect the choice of the accompanying article: the lighter the weight, the more easily *des* is used; the less light, the more *de* is likely. While the notion of grammatical weight is regarded as discrete in Abeillé and Godard (1999 and 2000), our notion of weight here is of a gradual nature and therefore can provide a relative characterization of every type (lexically) or token (in discourse) of adjectivals. It will be concluded that *petit* is located on the lowest end of the lexical scale of adjectives and tends to appear after *des*, which marks the "lightness" of the following adjective.

Résumé

En analysant les corpus de très grande taille, nous observerons d'abord quelques faits qui n'ont guère été abordés dans la littérature, notemment le degré de cooccurrence particulièrement fort entre *des* et *petit(e)s* dans le choix entre *de* et *des* devant les noms précédés d'épithète en français. Nous démontrerons ensuite, à partir d'études statistiques, que les critères explicatifs traditionnels que sont le niveau de langue et le degré de collocation entre l'adjectif et le nom, ne suffisent pas pour expliquer ces phénomènes. Nous proposerons en revanche deux autres facteurs, l'un étant la caractéristique lexicale des adjectifs, l'autre, la nature sémantico-fonctionnelle de l'épithète dans le discours. Ces facteurs constituent le degré de « poids » ou le degré d'importance informative de l'adjectif dans l'usage ; plus l'épithète est « légère », plus *des* est choisi, et moins elle l'est, plus *de* est sélectionné. Alors que la notion de « poids grammatical » proposée dans Abeillé et Godard (1999 et 2000) est de nature discontinue, la nôtre se place sur une échelle continue. C'est la méthode même de l'étude statistique qui nous oblige d'avancer cette dernière caractérisation. Avec cette caractérisation de « poids », nous pouvons traiter les facteurs lexicaux et discursifs de la même manière. Nous conclurons que *petit* est situé à l'extrémité de l'échelle lexicale et tend à apparaître précédé de *des*, qui fonctionne comme marqueur de la « légèreté » de l'épithète qui le suit.

Mots-clés : petit, adjectif épithète, poids, degré de collocation, français, lexical, discursif, article, weight

1. Introduction

Lorsqu'un adjectif au pluriel est antéposé par rapport au substantif, on a le choix en français entre *de* et *des* comme article indéfini, par exemple, *de bonnes conditions* et *des petits chiens*. Nous avons étudié statistiquement ce phénomène dans Fujimura *et al.* (2004 à paraître), du

point de vue aussi bien historique que stylistique en recourant à des corpus de très grande taille, et proposé plusieurs conditions qui déterminent cette variation.

Nous y avons non seulement constaté le facteur du niveau de langue et la question des mots composés, qui ont été répétés à maintes reprises dans la littérature relative à cette alternance, mais aussi découvert de nouveaux facteurs tels que la liaison phonétique : s'il y a une liaison entre l'adjectif et le nom, on note une tendance claire à éviter *des*. Dans cet article, nous allons examiner deux autres conditions, l'une étant la distribution disparate de *de* et de *des* suivant les adjectifs, et l'autre, la quantité d'informations que l'adjectif antéposé transmet dans le discours. Nous nous concentrerons surtout sur la question de *petit* qui présente une affinité extrêmement forte avec *des* par rapport à d'autres adjectifs.

La question du choix de l'article *de* ou *des* devant l'adjectif a un rapport étroit avec la tendance générale à l'antéposition ou à la postposition de cet adjectif par rapport au nom. Togeby dit par exemple : « Seulement une dizaine d'adjectifs courts et fréquents sont régulièrement antéposés dans la langue de la conversation, qui les fait accompagner alors de l'article partitif complet (= l'article *des*) : *C'est des bonnes nouvelles*,.. » (Togeby, 1982). Or, dans la littérature récente de la linguistique générale, la notion de « poids grammatical » a été proposée pour résoudre entre autres la question de la position de l'épithète en français (Abeillé et Godard, 1999 et 2000) ; l'adjectif est « léger » s'il est constamment antéposé au nom, tandis qu'il est « non-léger », s'il est toujours postposé. On pourrait donc supposer, en leur empruntant ces termes fort convenables, que *des* a tendance à être choisi si l'adjectif est « léger » et que *de* est sélectionné quand l'adjectif est « non-léger ». D'après les auteurs de ces termes, les adjectifs « légers » et les « non-légers » constituent deux catégories lexicales distinctes et discontinues. Pour nous, par contre, la notion de « poids » devrait être sur une échelle graduelle et continue. En définissant le « poids » comme « importance informative » ou « saillance informative », nous démontrerons que *petit* est l'adjectif le plus « léger » parmi tous les adjectifs français.

2. Adjectifs et *de* vs *des*

2.1. *Aperçu général*

Nous voyons dans la figure 1 les occurrences en nombre réel de *de* et de *des* qui précèdent les « ADJ + NOM ». Les données ont été recueillies à partir des corpus de la 2e moitié du 20e siècle (71,6 millions de mots au total), que nous présenterons ci-après.

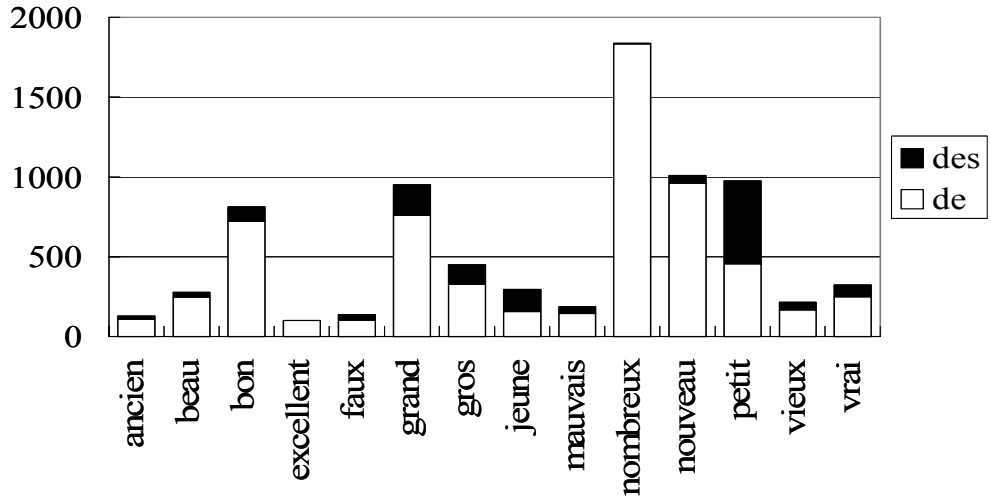


Figure 1. De vs des et adjectifs dans les textes contemporain

Le taux de *des* pour chaque adjectif est : 53.3% (*petit*), 46.4% (*jeune*), 27.2% (*gros)*, 24.8% (*faux)*, 23.1% (*vrai)*, 23.0% (*mauvais*), 22.7% (*vieux)*, 20.2% (*grand)*, 14.8% (*ancien)*, 11.2% (*bon)*, 10.5% (*beau)*, 4.8% (*nouveau)*, 0.2% (*nombreux)*, 0% (*excellent*) . Les « *de/des* + ADJ + NOM » les plus fréquents sont : *des petits problèmes, des jeunes filles, de gros problèmes, de faux papiers, de vrais morceaux, de mauvaises conditions, de vieux journaux, de grandes chances, d'anciens membres, de bonnes conditions, de beaux yeux, de nouvelles aventures, de nombreux cas* et *d'excellentes conditions*. On y trouve une disparité importante d'un adjectif à l'autre. La tendance de cooccurrence entre *des* et *petit* est remarquablement forte, tandis que *des* n'est guère compatible avec *nouveau*, *nombreux* et *excellent*. D'autres adjectifs se situent entre les deux. Quant à *jeune* qui est autant utilisé avec *des* que *petit*, nous le mettons de côté parce que c'est un adjectif qui ne qualifie en principe que les animés avec un pouvoir combinatoire limité.

Cette hétérogénéité de répartition de *de* et *des* demande une explication. Il y avait peu de grammairiens et linguistes qui l'avaient envisagée, ce phénomène n'ayant même pas été bien observé auparavant.

2.2. *Corpus et données*

Pour mener à bien cette étude, nous avons établi d'abord une base de données constituée d'environ 20 000 exemples avec « *de /des* + (ADV) + ADJ + NOM au pluriel » recueillis dans les corpus mentionnés dans le tableau 1 ci-après. Ce sont toujours des séquences avec une trentaine d'adjectifs tels que *beau, bon, grand, nouveau, petit, excellent etc.,* qui se trouvaient à suivre directement un lemme des verbes : *avoir*, *être* et *prendre*, et les prépositions : *à, avec, dans, par, pour* et *sur.* Nous y avons aussi inclus des exemples qui étaient sujets des verbes : *avoir, être* et *prendre*. Nous avons manuellement examiné tous les exemples pour ne conserver que ceux qui sont appropriés à notre recherche. Nous avons ainsi exclu tous les usages de *de* induits de la négation (ex. *on n'a pas eu de grandes discussions sérieuses)*, tous les homographes (ex. *la page de la liste odésienne est sur des nouvelles intéressantes*), toutes les occurrences de *de* ou *des* en tant que préposition (+ article) (ex. *les pièces sont de petites dimensions ; pour ce qui est des belles routes, ...*) etc.. L'examen des données nous enseigne que la fonction grammaticale du SN dans la phrase ne joue pas un rôle pertinent pour le choix entre *de* et *des*. Bien que notre base de données ne représente qu'une sous-classe du phénomène, cela n'entraîne pas de conséquence erronée pour notre but.

Période	Genre	Détail	Année	Nombre de mots (milliers)
20e s-2	Hansard	Les interventions en français dans le débat au Parlement Canadien, parlé officiel	1986-1988	3 300
	journal	*Le Monde* (Wordbanks *Online*)	1997 et 2001	13 000
		Libération (Wordbanks *Online*)	1992-1993	1 500
	revue	*Actuel* (Wordbanks *Online*)	1990-1992	2 000
		Marie Claire (Wordbanks *Online*)	1990-1993	3 100
	Forum de discussions (FD)	323 Forums de discussions abonnés, écrit non officiel sur les réseaux électroniques	10/2000	21 200
	roman	Frantext base catégorisée	1951-2000	15 600
	traité ou essai			11 900

20e s-1	roman	Frantext base catégorisée	1901-1950	20 600
	traité ou essai			13 500
19e s-2	roman	Frantext base catégorisée	1851-1900	14 500
	traité ou essai			4 400
19e s-1	roman	Frantext base non-catégorisée	1801-1850	12 600
	traité ou essai			8 100
18e s	roman	Frantext base non-catégorisée	1701-1800	12 000
	traité ou essai			11 200
17e s	roman	Frantext base non-catégorisée	1601-1700	3 800
	traité ou essai			7 000
		total		179 300

Tableau 1. Corpus de base

3. *Petit* et « poids lexical »

3.1. Petit *à travers l'histoire et les genres*

Dans cette section, nous allons examiner *petit* de plus près. Nous montrons tout d'abord avec le figure 2 que *petit* garde la spécificité d'être plus étroitement lié avec *des* que les autres adjectifs : *grand, bon, beau, nouveau,* dans toutes les périodes et dans tous les genres depuis le début de l'histoire de ce phénomène. Dans les figures, la ligne indiquant *petit* est constamment au-dessus des autres lignes, c'est-à-dire que *petit* a plus d'affinité avec *des* que les autres depuis toujours. Cette remarquable caractéristique de *petit* mérite de sérieuses recherches.

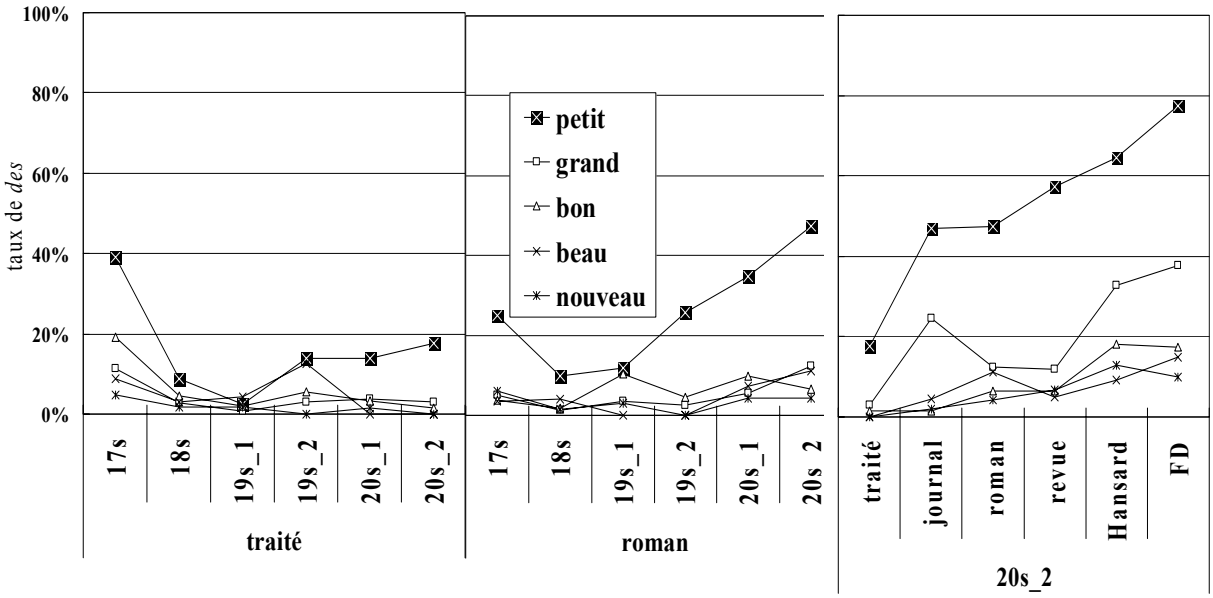


Figure 2. Taux de des *et adjectifs dans toute l'histoire et dans tous les genres*

3.2. Petit *et degré de collocation*

Togeby est un rare linguiste qui a fait remarquer cette forte tendance de cooccurrence, en disant que « petit *forme souvent avec le substantif une sorte de mot composé* » (1982 : 52). Nos données n'ont cependant pas relevé la spécificité de *petit* en tant que base de mots composés.

Nous avons créé, dans le but de l'étude des mots composés, une deuxième base de données constituée d'environ 3500 bigrammes (= 15000 tokens). Ce sont des « *ancien, beau, bon, grand, nouveau* et *petit* + NOM au pluriel » précédés de « (*à*| *avec*| *dans*| *derrière*| *devant*| *entre*| *malgré*| *par*| *parmi*| *pour*| *selon*| *sur*| *suivant*| *sous*) + (*de* | *des*) », recueillis à partir des périodiques contemporains d'environ 277 millions de mots : *Libération* (1995-2000)*, Le Monde* (1987-1992)*, Le Point* (1995-2000)*, Le Monde Diplomatique* (1984-1998) et *La Tribune* (1995-1999) distribués par CDROM-SNi. Nous avons choisi ces textes afin d'obtenir des données statistiques fiables. Ils sont aussi homogènes que gigantesques et faciles à traiter contrairement aux corpus en ligne tels que Frantext. Étant donné que ces textes ne sont pas étiquetés et que l'usage des analyseurs comme Cordial, FDG (Functional Dependency Grammar), TreeTagger ne nous semblait pas, après essai, avantageux pour notre but, nous avons restreint le contexte pour identifier le plus exactement possible les séquences qui correspondent à « ART + ADJ + NOM au pluriel ».

Afin de faire une comparaison avec leur degré d'affinité avec *des* (= taux de *des),* nous avons donné à chaque bigramme les informations concernant le degré de collocation entre l'adjectif et le nom : l'Information Mutuelle, le t-score et le z-score, calculées sous Excel par nous-mêmes (cf. Barnbrook, 1996 ; Oaks, 1998). Nous ne savions pas si le statut de mot composé était mesurable avec ces scores, mais sans aucun bon dictionnaire disponible définissant des mots composés français, nous n'avions pas d'autre moyen pour vérifier notre thèse. Les bigrammes avec les scores de collocation les plus élevés étaient : *petits boulots, bons offices, petits riens, belles empoignades, bonnes volontés* (IM)*, grandes entreprises, grands groupes, nouvelles technologies, grandes villes, grandes lignes* (t-score)*, petits boulots, grandes surfaces, anciens combattants, nouvelles technologies, bons offices* (z-score)*.* Voici les expressions développées avec lesquelles nous avons calculé les scores :

$$IM(x,y) = \log_2 \frac{F(x,y) * \text{Nombre total de mots}}{F(x)\,F(y)}$$

$$t\text{-score}(x,y) = \frac{F(x,y) - \frac{F(x)F(y)}{\text{Nombre total de mots}}}{\sqrt{F(x,y)}}$$

$$z\text{-score}(x,y) = \frac{F(x,y) - \frac{F(x)F(y)}{\text{Nombre total de mots}}}{\sqrt{F(x)F(y)}}$$

Nombre total de mots = 277 millions

Les résultats démontrent tout d'abord que les scores de degré de collocation entre l'adjectif et le nom sont en corrélation bien que faible avec le taux de *des* dans la totalité des bigrammes : les coefficients de corrélation de Pearson entre les taux de *des* et l'IM, le t-score et le z-score sont 0.255, 0.210, et 0.285 respectivement (n = 994). Pour ce calcul, nous avons pris en compte seuls les bigrammes qui apparaissaient plus de 2 fois dans l'environnement défini ci-dessus et, en même temps, plus de 9 fois au total dans toute circonstance dans le corpus : soit 994 types de bigramme (= 11301 tokens). On constate que le choix de *de* ou de *des* est conditionné par le degré de collocation entre l'adjectif et le nom, comme toutes les grammaires l'avaient indiqué à l'unanimité.

Ces scores n'expliquent cependant pas la diversité entre les adjectifs quant au choix de *de* ou de *des*, ainsi que le montre la figure 3. Le degré d'affinité entre l'adjectif et *des* indiqué par la ligne « taux de *des* » n'a de relation avec aucun des scores de collocation, ni avec l'IM, ni avec le t-score, ni avec le z-score. L'affinité forte entre *des* et *petit* n'est pas attribuable au figement étroit entre *petit* et le nom qui le suit.

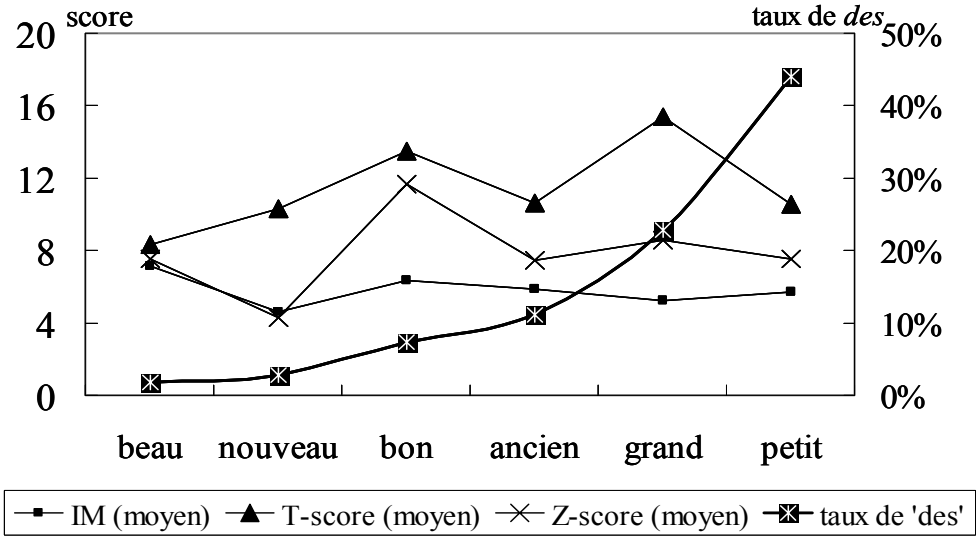


Figure 3. Taux de des*, adjectifs et scores de collocation*

Des a une très forte affinité avec *petit* dans toutes les périodes et dans tous les genres. Or, le degré de collocation entre l'adjectif et le nom est en corrélation avec le taux de *des*. Nos données ne relèvent pas néanmoins la particularité de *petit* en tant que base de mots composés.

3.3. *Contraste entre* petit *et* grand

Pour confirmer avec plus de sûreté que la question de *petit* est bien attribuable à sa propre nature et non à l'environnement où cet adjectif apparaît, nous avons ensuite observé le phénomène dans des conditions plus restreintes. Nous avons étudié, dans les corpus périodiques, l'occurrence de *de* et de *des* à côté des deux adjectifs antonymes : *petit* et *grand*, suivis d'un de ces noms : *actionnaires, affaires, appartements, banques, cahiers, centres, chefs, choses, compagnies, écrans, éditeurs, entreprises, établissements, films, firmes, fonds, groupes, hôtels, investisseurs, lunettes, maisons, marchés, moyens, opérations, pays, programmes, projets, quantités, restaurants, rôles, sacs, salles, séries, sociétés, structures, surfaces, tables, tentes, travaux, usines, valeurs* et *villes*. Il s'agit des noms qui peuvent se combiner librement avec *petit* et *grand* et qui sont effectivement apparus plus de 3 fois, associés avec chacun de ces deux adjectifs dans nos corpus.

Le résultat indique que même dans ces conditions limitées, *petit* présente une tendance forte de cooccurrence avec *des*, contrairement à *grand* qui a une tendance de cooccurrence avec *de* ($\chi 2 = 11.038$, dl = 1, $p < .01$).

	de	*des*
Grand	712	302
Petit	366	223

Tableau 2. De *vs* des *et* petit *vs* grand *(occurrences en nombre réel)*

Voici un exemple qui représente bien le contraste entre *petit* et *grand* :

(1) Or, cette relation entre la taille et le taux de mortalité est probablement un simple biais statistique, lié au fait qu'un taux calculé sur des petits effectifs est beaucoup plus fluctuant que quand il est calculé sur de grands effectifs. (*Le Monde*, 06/10/97)

et quelques autres auxquels on va revenir un peu plus bas :

(2) Pour tenter d'accroître ses parts de marché, la firme de Cupertino a signé depuis le début de l'année des accords de licence avec des petites sociétés comme Power Computing, Radius, ainsi qu'avec le japonais Bandaï. (*La Tribune*, 07/09/95)

(3) Il va y parvenir de façon spectaculaire: en quelques mois, il passe un accord de licence avec RSA Data Security - une firme spécialisée dans l'encryptage -, jette les bases d'un standard pour les serveurs sécurisés, fait considérablement évoluer le langage HTML (2) du Web et, surtout, signe des accords avec de grandes sociétés comme Bank of America, le groupe MasterCard, ainsi que le consortium CommerceNet qui regroupe les grands noms de l'industrie informatique. (*Libération*, 21/4/95)

3.4. *Caractéristique lexicale de* petit : *légèreté*

Nous avons démontré que la forte tendance de cooccurrence entre *des* et *petit* est seulement explicable avec une spécificité lexicale de cet adjectif. Ce phénomène n'est pas abordable en recourant à d'autres principes linguistiques, mais il est immanent au mot même de *petit*, qui est, d'après nous, l'adjectif « le plus léger » parmi ceux que nous avons examinés.

Cette caractéristique de *petit* n'est pas sporadique. Nous pouvons en effet trouver, autour de *petit,* d'autres phénomènes qui s'accordent bien avec la notion de « poids léger » définie comme « importance informative ». Nous donnons ici, évoqués par Fujita (2001), trois arguments qui soutiennent notre thèse que *petit* est un adjectif « très léger ».

3.4.1. *Unidimensionnalité*

Le premier est la question de l'unidimensionnalité de l'échelle *petit* / *grand* proposée dans Rivara, 1993. Selon l'auteur, les adjectifs antonymiques *bon / mauvais, beau / laid, nouveau / vieux* etc. sont bipolaires en ayant deux échelles indépendantes l'une de l'autre, alors que *petit* et *grand* portent une même échelle en commun : celle de grandeur. Sur cette échelle, *petit* est orienté vers le néant, tandis que *grand* l'est vers l'infini. On peut facilement dire que le pouvoir descriptif et qualificatif de *petit* est souvent restreint par rapport à celui de *grand* comme on peut l'observer dans les exemples 2 et 3 ci-dessus. L'adjectif de l'exemple 3 serait paraphrasable en *important* ou *extraordinaire,* par contre, celui de l'exemple 2 signifierait *peu important* ou *ordinaire.*

3.4.2. *Diminutif*

Le deuxième argument concerne le diminutif. En français, *petit* a souvent une valeur hypocoristique qui est en général exprimée dans d'autres langues par un suffixe ajouté à un mot (Herisson, 1956 ; Delhay, 1996). L'épithète *petit* dans cet usage n'implique aucune référence à la taille (ex. *Bonjour à ta petite famille)*. On peut dire qu'il s'agit d'une déficience sémantique, qui devrait aboutir à une dégradation syntaxique. Si ce genre de changement est surtout permis à *petit* et non à d'autres adjectifs, ce serait parce que le mot *petit* a une propriété adéquate à cette déviation, qui est d'après nous le caractère d'être « léger ».

3.4.3. Le moindre petit + *NOM*

Le troisième argument repose sur l'expression pléonastique : « *le moindre petit* NOM ». Cette tournure est fréquente et productive en français ; on a relevé 91 cas du « *le moindre petit* NOM » contre 2824 « *le moindre* NP » dans le FD (ex. *Je n'ai pas la moindre petite information à vous donner !* (FD, 04/09/2000)). Bien qu'il y ait des cas où le « *petit* NOM » dans la tournure constitue un mot composé (ex. *..dans la société du XIX*^{e} *siècle où le moindre petit bourgeois a au moins une bonne..*), ce n'est pas le cas de tous les exemples. Il nous semble que *petit* y perd son « poids » jusqu'à ce que la suite *moindre petit* ne soit plus redondante à l'oreille. A l'opposé, on n'entend jamais dire « *le meilleur bon* NOM », ni « *le pire mauvais* NOM », excepté dans les cas de mots composés comme *la meilleure bonne foi*.

On peut expliquer ce phénomène de la même manière que les faits observés ci-dessus. Le « poids » de *petit* étant très faible, cet adjectif devient donc non encombrant, il est facile de ne pas tenir compte de son existence sémantique et formelle. La comparaison des exemples suivants démontre bien que l'information transmise par le mot *petit* est minimale.

(4) Dans 99 % des cas, explique-t-on, c'est une fausse alerte. Mais les gens ne veulent pas prendre le moindre petit risque, car on n'en revient pas si le 1 % qui reste n'est pas faux. (*Le Monde*, 25/09/97)

(5) Les banques ne veulent plus prendre le moindre risque pour l'argent de leurs clients et exigent systématiquement le cautionnement. (*Le Monde*, 26/11/97)

4. « Poids discursif » ou quantité d'informations de l'épithète

Après avoir examiné le « poids lexical » de *petit* parmi d'autres adjectifs, nous allons passer à l'examen de la relation entre le « poids discursif » des épithètes et le choix de l'article *de* ou *des.* Notre hypothèse est la suivante : plus l'information que l'épithète véhicule est abondante, plus *de* est choisi ; moins elle l'est, plus *des* est sélectionné. L'épithète dans les mots composés se situe à une extrémité de cette échelle, car elle n'a aucune information à transmettre indépendamment du nom.

4.1. *Description des grammairiens*

Cette caractérisation de *de* et de *des* n'est pas une nouveauté. Au contraire, c'était une vue répétée dans la littérature, comme l'indiquent les citations suivantes.

Elle (=la langue) différenciait ainsi la qualification préalable, notoire, lexicalisée, qu'exprime parfois l'épithète antéposée (*des grands frères, des grandes sœurs, des petits pains,* etc.) de la qualification nouvelle, transitoire, voire prédicative, qui est son rôle dans bien des cas (*Elle a de grands yeux*, c'est-à-dire : Elle a les yeux grands). (*GLLF :* 260)

Sans l'article (= *de*), l'épithète a toute sa valeur propre, toute sa force de caractérisation accusée : *de beaux pays, de grands chirurgiens*. Avec l'article (= *des*) (...), la caractérisation s'efface un peu, ... (Le Bidois et Le Bidois, 1967 : 85).

Notre apport original consiste dans le point de vue de la gradualité. Cette notion permet de réunir divers facteurs sur une même échelle même s'ils paraissent hétérogènes à première vue.

4.2. *Centre d'information*

Le centre d'information est selon nous une autre manifestation du « poids non-léger ». Dans l'exemple suivant, où *de* est exceptionnellement employé devant *petits* dans le texte oral (1 cas de *de* contre 30 cas de *des* dans ce corpus du français parlé), *petits* joue le rôle de centre d'information dans le SN : *petits libraires*. Le mot *libraires* est dans ce dernier une répétition, à savoir une information superflue puisqu'il a déjà été dit en tant que sujet de la phrase copulative :

(6) les FNAC n'existaient pas les choses comme ça bon les l- les libraires étaient de petits libraires indépendants elle existe d'ailleurs toujours je pense la maison Gibert non (Giron S., 2001, Corpus Allier, PHYMO~54 (6,10 - 6,12))

Nous avons également attesté la préférence de *de* dans cet environnement : « NOM*i être* (*de* /*des*) ADJ NOM*j (i = j)* » dans les corpus du tableau 1 par exemple dans :

(7) Les artistes vivent souvent dans l'irréalité et à cause de cela, sont de grands artistes. (*Marie-Claire*)

(8) Si on considère que plus de 90 p. 100 des entreprises au Canada sont de petites entreprises, (Hansard)

(9) ... ramasser les légumes, aller à l' herbe pour les lapins, nourrir les poules..... les maisons sont de petites maisons comme la mienne, ...(Triolet E. *Le premier accroc coûte deux cents francs*)

4.3. *Renforcement adverbial*

Nous traitons en dernier lieu le renforcement de l'épithète par un adverbe. On peut facilement postuler que la présence d'un adverbe ajoute une information de plus ; on peut aussi estimer qu'elle rend la phrase plus complexe ou « plus lourde », en empruntant ce dernier terme à Arnold et al., 2000. L'épithète devenant « moins légère », la possibilité du choix de *de* augmente.

Cette tendance est clairement démontrée dans le tableau 3, qui relève les occurrences de *de* et de *des* devant les adverbes *fort, moins, plus, si, tout, très et trop* dans les corpus : 2ème moitié du 20e siècle, présentés dans le tableau 1. S'il y a un adverbe entre l'article et l'adjectif, *des* n'est guère sélectionné, mais *de* est de fait presque toujours choisi ($\chi 2 = 38.4887$, dl = 1, $p < .01$).

	de	*des*		*de*	*des*		*de*	*des*
sans adverbe	5503	1122	*tout / toutes*	22	3	*fort*	7	0
avec adverbe	251	6	*plus*	21	2	*trop*	6	0
très	173	1	*si*	17	0	*moins*	5	0

Tableau 3 : De *vs* des *et adverbe*

Cette tendance n'est cependant pas non plus une règle absolue. Alors que l'usage de *des* est rare avec *tout / toutes*, il n'est pas aussi exceptionnel qu'avec *très*. On peut en effet noter une différence statistiquement significative de l'occurrence de *des* avec *tout / toutes* et avec *très* ($\chi 2 = 14.4871$, dl = 1, $p < .01$). Les exemples suivants relèvent le contraste de *très* et *tout / toutes* : « *des tout / toutes* ADJ + NOM » est employé même dans le corpus écrit académique dans lequel l'usage de *des* est généralement limité comme dans (10), tandis que « *de très* ADJ + NOM » est utilisé dans le corpus parlé où l'usage de *des* est en général fréquent comme dans (11).

(10) Elle (= la diarrhée) apparaît dans un élevage sur des tout jeunes poussins de trois à six jours. (Garcin E. *Guide vétérinaire*)

(11) hein quand tu travailles comme ça sur les marchés tu as les tu as de très bonnes relations avec eux et puis voilà donc je crois que c'est ça qu'il apprécie voilà mmh (Giron S., 2001, Corpus Allier, FRUIT~22 (13,16 - 14,2))

Or, on sait bien qu'il y a une disparité quant au choix de l'adverbe selon les adjectifs. *Très* se combine avec des adjectifs qualificatifs typiques, comme *grand, beau, bon,* alors que *tout* s'associe, d'un emploi beaucoup plus limité, avec des adjectifs qui ont « une limite idéale » (Hanse, 1987 : 950) comme ceux de couleur : *tout rouge*, par exemple. *Tout* est un adverbe plus notionnel et plus logique que *très,* qui est plutôt emphatique, marquant sans réserve un degré très élevé. Nous pouvons penser que cette différence sémantico-fonctionnelle est la raison pour laquelle « *des tout* ADJ NOM » est moins exceptionnel que « des très ADJ NOM » dans le corpus. Le rôle de *de* est de mettre en relief la présence d'un adjectif « non léger » qui a un caractère typiquement qualificatif dans la position qui le suit.

D'après nous, *tout* est un adverbe « plus léger » que *très,* contrairement à ce qu'ont proposé Abeillé et Godard (1999 et 2000). Selon ces auteurs, ce sont tous les deux des adverbes de la catégorie du « poids léger » au même titre que *trop, assez, vraiment, peu*, à l'opposé de *politiquement, véritablement* ou *absolument* qui sont « non-légers ». Le critère pour cette classification réside dans la position de l'épithète : « adverbe + adjectif » par rapport au nom. Avec les adverbes « légers », l'adjectif peut rester antéposé comme dans « *de très bonnes conditions* », alors qu'avec les « non-légers », il doit être postposé comme dans « *une décision politiquement habile* ». Nous voudrions en revanche avancer pour notre part l'idée de gradualité du « poids », en recourant à la fréquence comme critère d'évaluation. Si le taux de *des* est plus élevé avec *tout* qu'avec *très* et que les autres conditions sont égales, c'est parce que *tout* est « plus léger » que *très.* Il en va de même pour les adjectifs comme nous l'avons déjà affirmé : *petit* est l'adjectif « le plus léger » parmi ceux que nous avons examinés. *Excellent* et *nombreux* sont par contre supposés « les moins légers » parmi eux, bien qu'ils soient eux-mêmes « plus légers » que ceux qui sont toujours postposés, comme *carré, français* ou *présidentiel*.

5. Conclusion

Partis de l'observation de l'adjectif *petit* qui a une forte tendance de cooccurrence avec l'article *des* depuis le 17^{e} siècle jusqu'à nos jours et à travers tous les genres de texte, en passant par la réfutation de l'argument « *petit* comme base de mots composés » au moyen de quelques techniques statistiques, nous avons abouti à l'idée de « poids », notion d'une portée importante qui permet d'expliquer de nombreux facteurs pertinents de ce phénomène.

Nous avons proposé dans cet article deux sous-classes de « poids ». L'un est le « poids lexical » et l'autre, le « poids discursif ». Le premier concerne un caractère immanent du mot,

tandis que l'autre repose sur le rôle sémantico-fonctionnel de l'expression dans le discours. Voici les échelles relevées comme pertinentes pour rendre compte du phénomène :

poids de l'épithète

des **<=> plus léger < ---------- > moins léger <=>** ***de***

lexical

petit > .. grand ..beau, .. nouveau > .. nombreux, excellent > .....(carré, présidentiel, ..)

tout > très

discursif

moins informatif > plus informatif

constituant d'un mot composé > épithète pleine

sans adverbe > avec adverbe

Bien que l'idée de « poids » nous ait été directement inspirée par Abeillé et Godard (1999 et 2000), la différence entre notre point de vue et le leur est claire. Pour eux, le « poids » est une notion qui sert à classer des éléments dans des catégories discontinues, tandis que pour nous c'est une échelle continue et graduelle. On peut dire que c'est la méthode même de l'étude statistique qui nous oblige à avancer cette dernière caractérisation du « poids ». Cette position n'est d'ailleurs pas vraiment nouvelle, au contraire, mais bien orthodoxe dans la littérature d'une longue tradition concernant le « poids » et l'ordre des mots (cf. Arnold *et al.*, 2000 ; Wasow, 1997 entre autres). Une de nos contributions y serait l'application de cette notion dans un autre domaine que l'ordre des mots.

Nous voudrions enfin signaler une possibilité de rendre compte d'un autre facteur qui détermine le choix entre *de* et *des* : la liaison phonique, avec la notion de « poids ». S'il y a une liaison entre l'épithète et le nom, *de* est préféré, sinon, *des* l'est. Le « poids » concerné pourrait être phonétique, parce qu'avec une liaison, le mot devient plus long. Ou bien, il pourrait porter sur la quantité d'information, parce qu'avec la liaison, une information concernant la pluralité est transmise en plus. Voilà la question intéressante qui demande une exploitation.

Références

Abeillé A. et Godard D. (1999). La place de l'adjectif épithète en français : le poids des mots. *Recherches Linguistiques*, vol. (28) : 9-31.

Abeillé A. et Godard D. (2000). French Word Order and Lexical Weight. In Borsley R. (Ed.), *The Nature and Function of Syntactic Categories, Syntax and Semantics*, vol. (32) : 325-360.

Arnold J., Wasow Th., Losongco A. et Ginstrom R. (2000). Heaviness vs. Newness : The effects of complexity and information structure on constituent ordering. *Language*, vol. (76) : 28-55.

Barnbrook G. (1996). *Language and Computers: A Practical Introduction to the Computer Analysis of Language*. Edinburgh U. P.

Delhay C. (1996). *Il était un « petit X » : pour une approche nouvelle de la catégorisation dite diminutive*. Larousse.

Dupré (1972). *Encyclopédie du bon français dans l'usage conte*mporain, t. (II).Éditions de Trévise.

Fujimura I., Uchida M. et Nakao H. (2004). Opposition entre *de* et *des* devant les noms précédés d'épithète en français : étude descriptive. In *Actes des 3èmes journées de la linguistique du corpus*,. A paraître.

Fujita T. (2001). Tokushu na saijo-kyu *le moindre* (*Le moindre* : un superlatif atypique), *Grant-in-Aid for COE Research Report (5) : Researching and Verifying an Advanced Theory of Human Language*. Kanda University of International Studies : 177-195.

Giron S. (2001). Corpus Allier.

Grand Larousse de la Langue Française (1971), t. (1). Larousse.

Hanse J. (1987). *Nouveau dictionnaire des difficultés du français moderne*. Duculot.

Herisson Ch. D. (1956). Le diminutif hypocoristique « *petit* ». *Le Français Moderne,* vol. (XXIV) : 35-47.

Le Bidois G. et Le Bidois R. (1967). *Syntaxe du français moderne*, t. (1). Éd. A. Picard.

Togeby K. (1982). *Grammaire française, vol.1 : Le Nom*. Akademisk Forlag.

Oakes M.P. (1998). *Statistics for Corpus Linguistics.* Edinburgh U. P.

Rivara R. (1993). Adjectifs et structures sémantiques scalaires. *Information grammaticale*, vol. (58) : 44-56.

Wasow Th. (1997). Remarks on Grammatical Weight. *Language Variation and Change*, vol. (9) : 81-105.

EDITE MEDITE : un logiciel de comparaison de versions

Jean-Gabriel Ganascia[1], Irène Fenoglio[2], Jean-Louis Lebrave[2]

[1]LIP6 – Université Pierre et Marie Curie – 8 rue du Capitaine Scott – 75015 Paris – France
[2]ITEM – CNRS – 45 rue d'Ulm – 75005, Paris – France
Jean-Gabriel.Ganascia@lip6.fr

Abstract

MEDITE has been designed to facilitate, with a systematic quantification, the preliminary study required by any textual genetic interpretation and, more generally, to help the philologists. It is to compare two states of literary texts by pointing at the textual transformations between them. At the heart of the program, the main algorithm comprises three steps: the detection of the maximal disjoints common blocs, the identification of the pivots and shifts, and the computation of the deletions, insertions and replacements. Lastly, an interface renders the obtained results easily visible and makes the user able to take notes.

Résumé

MEDITE a été conçu pour faciliter, grâce à une quantification systématique, l'étude préalable à toute interprétation de génétique textuelle et, plus généralement, pour aider les philologues. Il s'agit de comparer deux états de textes littéraires en indiquant les transformations textuelles opérées de l'un à l'autre. Au cœur de ce programme, l'algorithme principal comprend trois phases : la détection des blocs communs maximaux disjoints, l'identification des pivots et des déplacements et le calcul des suppressions, des insertions et des remplacements. Enfin, une interface visualise les résultats obtenus et permet à l'utilisateur de prendre des notes.

Mots-clés : séquences, homologies, genèse textuelle, opérations linguistiques,

1. Introduction

Le travail d'ajustage auquel se livre l'écrivain dans ses repentirs, ses coupes ou ses insertions successifs fait ici l'objet d'une étude systématique à l'aide d'outils informatiques qui s'inspirent partiellement de ceux développés pour l'étude des macromolécules biologiques. Nous avons réalisé un programme qui repère automatiquement les opérations structurales qui font passer d'un texte à un autre. Ces transformations élémentaires (déplacements, insertions, suppressions et remplacements de blocs de caractères), identifiées depuis longtemps par les spécialistes de la génétique textuelle (de Biasi, 2000 ; Hay, 2002 ; Contat et Ferrer 1998 ; Gresillon, 1994 ; Lebrave, 1984 et 1990 ; etc.), peuvent ensuite être associées aux catégories syntaxiques ou sémantiques des mots ou des groupes de mots pour donner naissance à des opérations linguistiques de réécriture (déplacement d'un adverbe, remplacement d'un mot par un hyperonyme ou par un hyponyme, suppression ou ajout d'un adjectif etc.) (Fenoglio et Boucheron, 2002). Notre logiciel mime les opérations exécutées à la main par le philologue qui compare des textes. Il comprend une interface permettant aussi bien de visualiser les modifications faisant passer d'un état du texte à un autre, que de recenser toutes les modifications, ajouts, suppressions, déplacements ou remplacements. L'automatisation autorise à la fois une répétition à l'identique de ces opérations, et une systématisation de la démarche sur des textes longs qu'il eut été très pénible de traiter manuellement. On peut ainsi travailler sur des articles, voire sur des livres entiers, et procéder à des études statistiques afin de caractéri-

ser le style de réécriture de tel ou tel auteur, et d'identifier, pour un même auteur, les différentes phases de réécriture : expansion, resserrement, …

De nombreuses applications sont envisagées. Originellement, le projet fut conçu pour la critique génétique : il s'agissait d'aider à comparer des brouillons d'auteurs, afin de saisir la nature du travail de réécriture. D'autres applications sont envisagées, en particulier la comparaison de variantes pour la littérature médiévale. (Cf. « Éloge de la variante » (Cerquiglini, 1989))

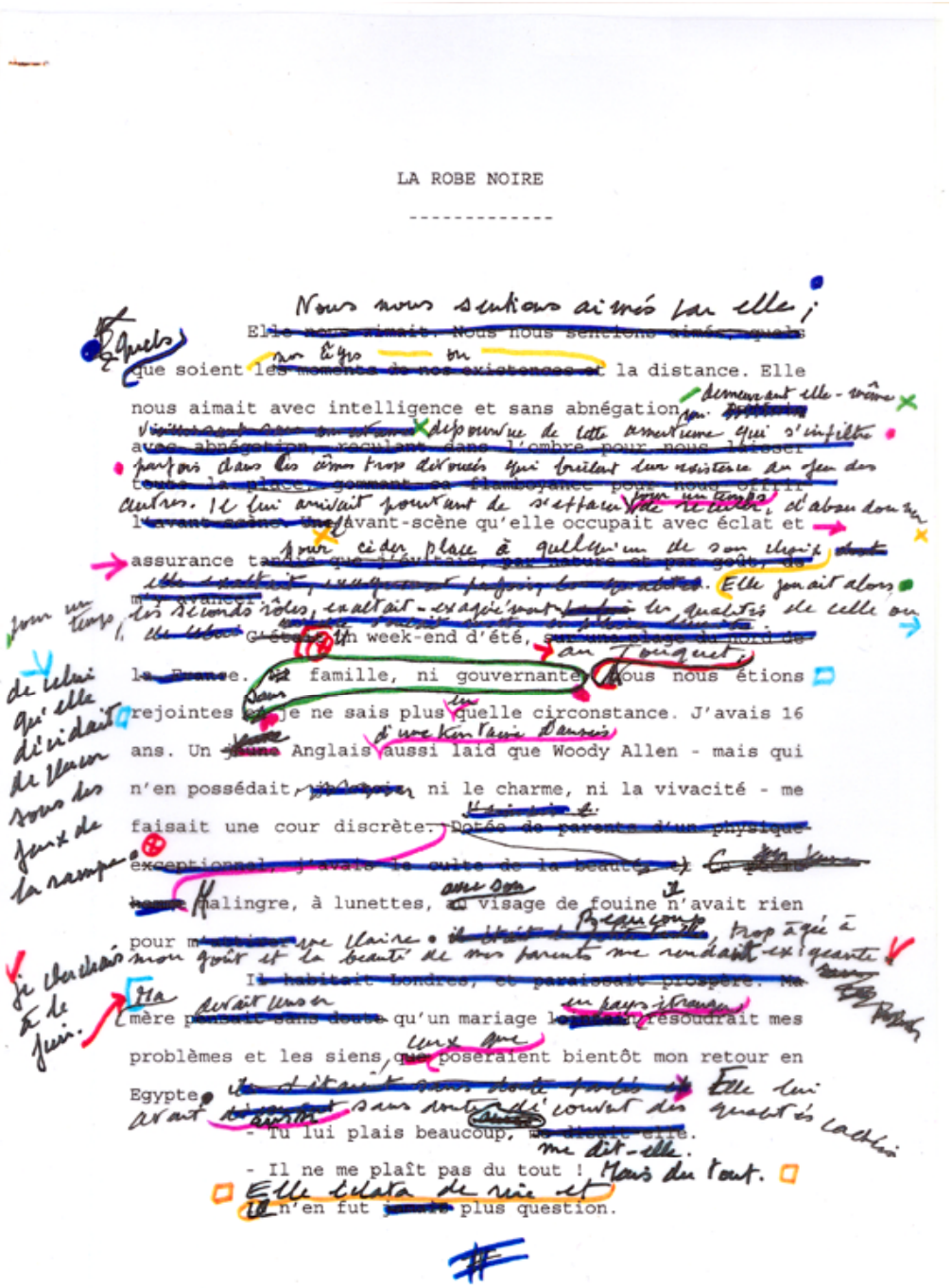


Figure 1. Une page d'un manuscrit de « La robe noire » (Andrée Chedid)

Cet article est consacré à la présentation du logiciel MEDITE, à ses fondements algorithmiques et à son interface de visualisation. Plus précisément, l'article se divise en quatre parties : après avoir précisé le sens d'un certain nombre de termes techniques employés par les généticiens du texte, nous aborderons les fondements algorithmiques du programme, puis, dans une troisième partie, nous présenterons l'interface graphique.

2. Signification de quelques termes techniques

La génétique textuelle étudie les processus d'écriture des textes à partir des traces. Généralement, du moins pour la plupart des auteurs d'avant l'âge informatique, des brouillons rassemblent ces traces sous forme soit intégralement manuscrite, soit partiellement manuscrite et partiellement tapée à la machine, soit totalement tapée à la machine.

À titre d'illustration, la figure 1 contient la photographie d'un brouillon d'auteur. En l'occurrence, le début d'une nouvelle d'Andrée Chedid, « La robe noire » paru dans le recueil Les saisons de passage (Chedid, 1996).

2.1. *Versions*

Dans la suite, nous distinguerons les différents supports matériels, c'est-à-dire les brouillons successifs, comme autant de *versions* de l'œuvre. Ainsi, dans le dossier génétique qui nous intéresse, celui du roman d'Andrée Chedid « La robe noire », l'auteur a recopié cinq fois son texte, donnant naissance à cinq versions.

2.2. *État*

Comme on peut le constater sur la figure 1, chaque brouillon est annoté, raturé, réécrit, ce qui rend la lecture assez confuse. Toutefois, les chercheurs savent identifier, avec une plus ou moins grande certitude, les différents états du texte, c'est-à-dire les différents textes présents sur une même version. Chacun de ces états correspond à une transcription linéaire qui fait abstraction de l'information visuelle et de la spatialisation : inscriptions marginales, couleurs, notes, etc. tout y est réduit à du texte brut. Sur notre exemple, l'état premier est identifié au texte tapé à la machine et les différentes couleurs du manuscrit sont associées aux différentes campagnes de réécriture et de relecture. Pour faciliter la présentation des choses, nous ne considérerons ici que deux états : le tapuscrit et l'état final (voir figure 2).

Bien évidemment, ce n'est là qu'un artifice de présentation. Il appartient dans chaque cas au chercheur de définir les différents états qu'il veut considérer.

Une fois ces états identifiés, le logiciel MEDITE va les comparer de façon à retrouver automatiquement les opérations de réécriture, pour en faire l'inventaire. Ceci étant, il convient de bien noter que la mise en œuvre du logiciel MEDITE présuppose qu'un travail préalable ait dégagé, à partir des différentes versions, les différents états du texte sous forme d'autant de transcriptions linéaires de ce même texte.

État initial : tapuscrit	État final
Elle nous aimait. Nous nous sentions aimés, quels que soient les moments nos existences et la distance. Parfois avec abnégation, reculant dans l'ombre pour nous laisser toute la place, gommant sa flamboyance pour nous offrir l'avant-scène. Une avant-scène qu'elle occupait avec éclat et assurance tandis que j'évitais, par	Nous nous sentions aimés par elle ; quels que soient nos âges ou la distance. Elle nous aimait avec intelligence et sans abnégation demeurant elle-même dépourvue de cette amertume qui s'infiltre parfois dans les âmes trop dévouées qui brûlent leur existence au feu des autres. Il lui arrivait pourtant de s'effacer pour un

nature et par goût, de m'y avancer. C'était un week-end d'été, sur une plage du nord de la France. Ni famille ni gouvernante, nous nous étions rejointes en je ne sais plus quelle circonstance. J'avais 16 ans. Un jeune Anglais aussi laid que Woody Allen - mais qui n'en possédait, je crois, ni le charme, ni la vivacité - me faisait une cour discrète. Dotée de parents d'un physique exceptionnel, j'avais le culte de la beauté et ce petit homme malingre, à lunettes, au visage de fouine n'avait rien pour m'attirer. Il habitait Londres, et paraissait prospère. Ma mère pensait sans doute qu'un mariage lointain résoudrait mes problèmes et les siens que poseraient bientôt mon retour en Egypte. - Tu lui plais beaucoup, me disait-elle. - Il ne me plaît pas du tout ! Il n'en fut jamais plus question.	temps, de reculer, d'abandonner l'avant-scène qu'elle occupait avec éclat et assurance pour céder place à quelqu'un de son choix. Elle jouait alors pour un temps, les seconds rôles, exaltait - exagérément parfois les qualités de celle ou de celui qu'elle décidait de placer sous les feux de la rampe. Un week-end d'été au Touquet, nous nous étions rejointes sans famille, ni gouvernante, je ne sais plus en quelle circonstance. J'avais 16 ans. Un jeune Anglais d'une trentaine d'années aussi laid que Woody Allen - mais qui n'en possédait ni le charme, ni la vivacité - me faisait une cour discrète. Malingre, à lunettes, avec son visage de fouine il n'avait rien pour me plaire. Beaucoup trop âgé à mon goût et la beauté de mes parents me rendait exigeante. Je cherchais à le fuir. Ma mère devait penser qu'un mariage en pays étranger résoudrait mes problèmes et les siens, ceux que poseraient bientôt mon retour en Egypte. Elle lui avait aussi sans doute découvert des qualités cachées. - Tu lui plais beaucoup, me dit-elle. - Il ne me plaît pas du tout ! Mais du tout. Elle éclata de rire et il n'en fut jamais plus question.

Figure 2. État initial et état final du texte d'Andrée Chedid dans le manuscrit de la figure 1

3. Fondements algorithmiques

Comme nous venons de le voir, le programme MEDITE prend en entrée deux états d'un même texte de façon à repérer les transformations qui font passer de l'un à l'autre, ou, plus exactement, l'ensemble minimal de transformations qui font passer du texte initial au texte « corrigée ». Formulé de la sorte, le problème apparaît très proche de celui posé par le calcul des « distances d'édition » (Sankoff et Kruskal, 1983 ; Crochemore et Rytter, 1994 ; Ganascia, 2001). Rappelons que la notion de « distance d'édition » se fonde sur des opérateurs de transformation, que l'on appelle en termes techniques des « éditions », car ils modifient des chaînes de caractères, et sur la minimisation du coût des transformations qui font passer d'une séquence à une autre.

Dans un premier temps, nous avions cru pouvoir réutiliser les distances d'édition, d'où l'acronyme du projet, EDITE, qui fait référence aux dites « éditions » et qui signifie « Étude Diachronique et Interprétative du Travail de l'Écrivain ». Or, il s'est rapidement avéré que cette utilisation des distances d'édition n'était pas possible, du moins telle quelle. En effet, il n'existe de procédure efficace de calcul de la distance d'édition que pour des ensembles d'éditions très restreints, comme l'ensemble dit standard qui comprend les trois opérations de *suppression*, d'*insertion* et de *remplacement*. Dans la mesure où la détection des *déplacements* joue un rôle important pour la génétique textuelle, et que l'introduction des *déplacements* dans l'ensemble des éditions change totalement la complexité algorithmique de la procédure de calcul des distances, il est nécessaire de procéder autrement. De plus, la taille des

textes (plusieurs centaines de milliers de caractères) interdit l'emploi de procédures d'une complexité polynomiale : il faut se limiter à une complexité linéaire ou, au plus, à une complexité en O(n*lg(n)), n étant la longueur des textes à traiter.

Afin de réduire la complexité et de répondre au mieux au problème posé, nous avons donc conçu un algorithme spécifique qui procède en trois étapes :

1. Détection des blocs communs maximaux disjoints ;
2. Identification des déplacements et des pivots ;
3. Calcul des insertions, des suppressions et des remplacements.

3.1. *Détection des blocs communs maximaux disjoints*

La détection des blocs maximaux fait appel à des algorithmes classiques (Karp, Miller et Rosenberg, 1972 ; Landraud, Avril et Chrétienne, 1989) de recherche d'homologies dans les séquences. Nous n'insisterons donc pas sur la mise en œuvre de ces algorithmes, sauf à dire qu'il y a parfois des recouvrements entre blocs maximaux. Ainsi, les deux chaînes « Il a avalé » et « Il avala » donnent deux blocs maximaux, |Il a | et | aval| qui se recouvrent partiellement. Pour obtenir des blocs maximaux disjoints, il faut introduire une césure. Or, généralement il y a plusieurs possibilités. Sur notre exemple, il y en a trois : |Il↑ a | et |↑ aval|, |Il ↑a | et | ↑aval| ou |Il a↑ | et | a↑val|, ce qui donne, en soulignant les insertions et les suppressions sur les deux chaînes initiales, les trois solutions suivantes : « Il| a| aval|é » et « Il| aval|a », « Il |a |aval|é » et « Il |aval|a » ou « Il a| a|val|é » et « Il a|val|a ». Dans la mesure du possible, il faut éviter la fragmentation des mots, c'est pourquoi nous avons choisi de mettre en priorité la césure sur les signes de ponctuation ou sur les blancs.

Par ailleurs, toujours pour éviter la fragmentation excessive des mots, nous ne mentionnons que les blocs communs d'une taille supérieure à une valeur seuil fixée arbitrairement. Par défaut ce seuil est de 4, ce qui veut dire qu'avant introduction de la césure, les homologies doivent avoir une longueur supérieure à 4 caractères. De la sorte, nous repérons des mots isolés de longueur supérieure à deux caractères, sachant qu'ils sont entourés de deux frontières de mots (blanc ou signe de ponctuation), ainsi que les préfixes ou les suffixes de plus de trois caractères, ce qui correspond à une syllabe.

Notons que la longueur minimale des blocs est un paramètre qui peut être modifié par l'utilisateur, sans difficulté (Voir section 4.4.3.). Cependant, du fait de l'introduction d'une césure qui supprime les recouvrements, il se peut que des blocs de longueur inférieure à la limite inférieure apparaissent dans les blocs communs. Cela signifie que ces blocs appartiennent à des blocs communs de longueur supérieure à la valeur seuil, mais qu'ils ont été rognés pour éviter des recouvrements.

3.2. *Identification des déplacements et des pivots*

Parmi l'ensemble des blocs communs maximaux disjoints, certains se retrouvent dans le même ordre dans les deux textes, le texte source et le texte corrigé, tandis que d'autres apparaissent déplacés. Ainsi, si nous avons la séquence de blocs maximaux B_1 B_2 B_3 B_4 B_5 dans le texte source et la séquence B_2 B_3 B_1 B_4 B_5 dans le texte corrigé, on peut inférer que le bloc B_1 a vraisemblablement été déplacé, même si cette appréciation est subjective, car on pourrait tout autant dire que ce sont les blocs B_2 et B_3 qui ont été déplacés. L'algorithme que nous avons mis en œuvre détermine les blocs déplacés en essayant de minimiser l'amplitude des déplacements mesurée en nombre de caractères. Plus exactement, cet algorithme prend en

considération la taille des blocs maximaux de façon à minimiser le nombre de déplacements de caractères requis pour passer d'une séquence de blocs maximaux à l'autre.

À l'issue de cette phase, on distingue parmi les blocs maximaux disjoints, des blocs dits « déplacés » et des blocs qui apparaissent dans le même ordre dans le texte source et dans le texte cible. Ces derniers sont appelés les « blocs pivots », ou plus simplement les « pivots » de la comparaison.

3.3. *Calcul des insertions, des suppressions et des remplacements*

Une fois déterminés les « blocs pivots » et les « blocs déplacés », il reste à calculer les *suppressions*, les *insertions* et les *remplacements*. Le programme procède comme suit :

- Lorsque deux pivots P et P' sont jointifs dans le texte source, la chaîne qui sépare P et P' dans le texte corrigé correspond à une *insertion*. Notons, pour éviter tout malentendu, que les deux pivots P et P' ne peuvent être jointifs à la fois dans le texte source et dans le texte corrigé, car sinon, P et P' ne serait maximaux ni l'un, ni l'autre.
- Lorsque deux pivots P et P' sont jointifs dans le texte corrigé, la chaîne qui sépare P et P' dans le texte source correspond à une suppression.
- Enfin, lorsque deux pivots P et P' ne sont jointifs ni dans le texte source, ni dans le texte corrigé, on dit qu'il y a remplacement de la chaîne comprise entre P et P' dans le texte source, par la chaîne comprise entre P et P' dans le texte corrigé.

À ce stade, il convient de préciser qu'un même bloc peut être à la fois déplacé et se trouver dans une insertion, une suppression ou un remplacement. En effet, dans le cas de déplacements de petits blocs situés à l'intérieur d'une insertion, d'une suppression ou d'un remplacement, il est souvent préférable de considérer que c'est l'ensemble qui est inséré, supprimé ou remplacé, de façon à éviter une fragmentation excessive des textes. Nous avons introduit deux paramètres qui permettent de lisser plus ou moins les résultats et d'inclure les blocs déplacés dans les insertions, les suppressions ou les remplacements.

Dans tous les cas, l'information sur le déplacement n'est pas omise. Elle vient se surajouter à d'autres informations. C'est particulièrement important pour repérer qu'un mot est « libéré » par un auteur afin d'être réemployé plus loin dans le même texte, sans commettre de répétition.

En conclusion, notons que notre algorithme a été testé sur de nombreux textes d'écrivains. En confrontant les résultats obtenus avec des interprétations philologiques, on constate que la plupart du temps, on retrouve les déplacements, les insertions, les suppressions et les remplacements déjà identifiés manuellement par les généticiens du texte.

4. Interface de visualisation

Pour faciliter la lecture des résultats, nous avons programmé une interface de visualisation qui comporte trois fenêtres (voir figure 3) : deux fenêtres dans la partie supérieure, l'une destinée au texte source, l'autre au texte corrigé, et une fenêtre dans la partie inférieure dont le contenu peut varier comme nous allons le voir ici.

4.1. *Partie supérieure : les textes*

Pour faire fonctionner MEDITE, il faut charger d'abord le texte source dans la fenêtre de gauche et le texte corrigé dans la fenêtre de droite, ce qui se fait, dans l'un et l'autre cas, comme dans un éditeur classique, avec des menus déroulant.

Un bouton permet ensuite de lancer la comparaison au moyen de l'algorithme précédemment décrit. Les résultats s'affichent alors en couleur : insertions, suppressions et remplacements sont marqués chacun par une couleur spécifique, que l'on peut faire varier à loisir. De plus, les blocs déplacés sont soulignés, ce qui autorise une superposition des deux indications : déplacements d'un côté, insertions, suppressions ou remplacements de l'autre.

Enfin, comme sur de très longs textes le lecteur est susceptible de se perdre, un compteur indique au-dessus de chacune des deux fenêtres de la partie supérieure, le numéro d'ordre du premier pivot présent dans la fenêtre de visualisation. L'utilisateur a alors tout loisir de faire défiler les textes à l'aide des ascenseurs, pour mettre les pivots de l'un et de l'autre texte en regard. Pour faciliter encore les choses, il est possible, en cliquant sur un pivot, de faire automatiquement défiler le texte homologue, dans l'autre fenêtre, jusqu'à ce que le pivot correspondant soit mis en regard du premier.

4.2. *Partie inférieure : information d'usage*

Le contenu de la fenêtre inférieure est spécifié au moyen des différents onglets qui apparaissent au bas de l'interface : TRANSFORMATIONS, COMMENTAIRES, LÉGENDE, PARAMÈTRES.

4.2.1. *Transformations*

Par défaut, l'onglet « TRANSFORMATIONS» est activé et la fenêtre contient l'ensemble des transformations qui font passer du texte source au texte corrigé, à savoir l'ensemble des insertions, des suppressions, des remplacements et des déplacements.

4.2.2. *Légende*

L'onglet « LÉGENDE » fait apparaître la légende de l'interface, c'est-à-dire la signification des couleurs, par exemple ici, bleu pour les insertions et les suppressions, vert pour les remplacements et souligné pour les déplacements. Au reste, il est loisible, de modifier manuellement les couleurs des insertions, des suppressions et des déplacements, ainsi que le style des déplacements.

4.2.3. *Paramètres*

Nous avons précédemment mentionné trois paramètres, l'un porte sur la taille minimale des blocs maximaux recensés, les deux autres, sur le lissage au cours du calcul des insertions, des suppressions et des remplacements. Ces trois paramètres sont accessibles dans la fenêtre du bas, à l'aide de l'onglet « PARAMÈTRE». On se trouve alors en mesure de modifier ces paramètres à volonté.

4.2.4. *Commentaires*

Enfin, l'onglet « COMMENTAIRES» fait apparaître une fenêtre vide où il est possible d'insérer des notes réutilisables par la suite. De même, on peut coller des parties du texte, ou des transformations, de façon à préparer un article ou une analyse.

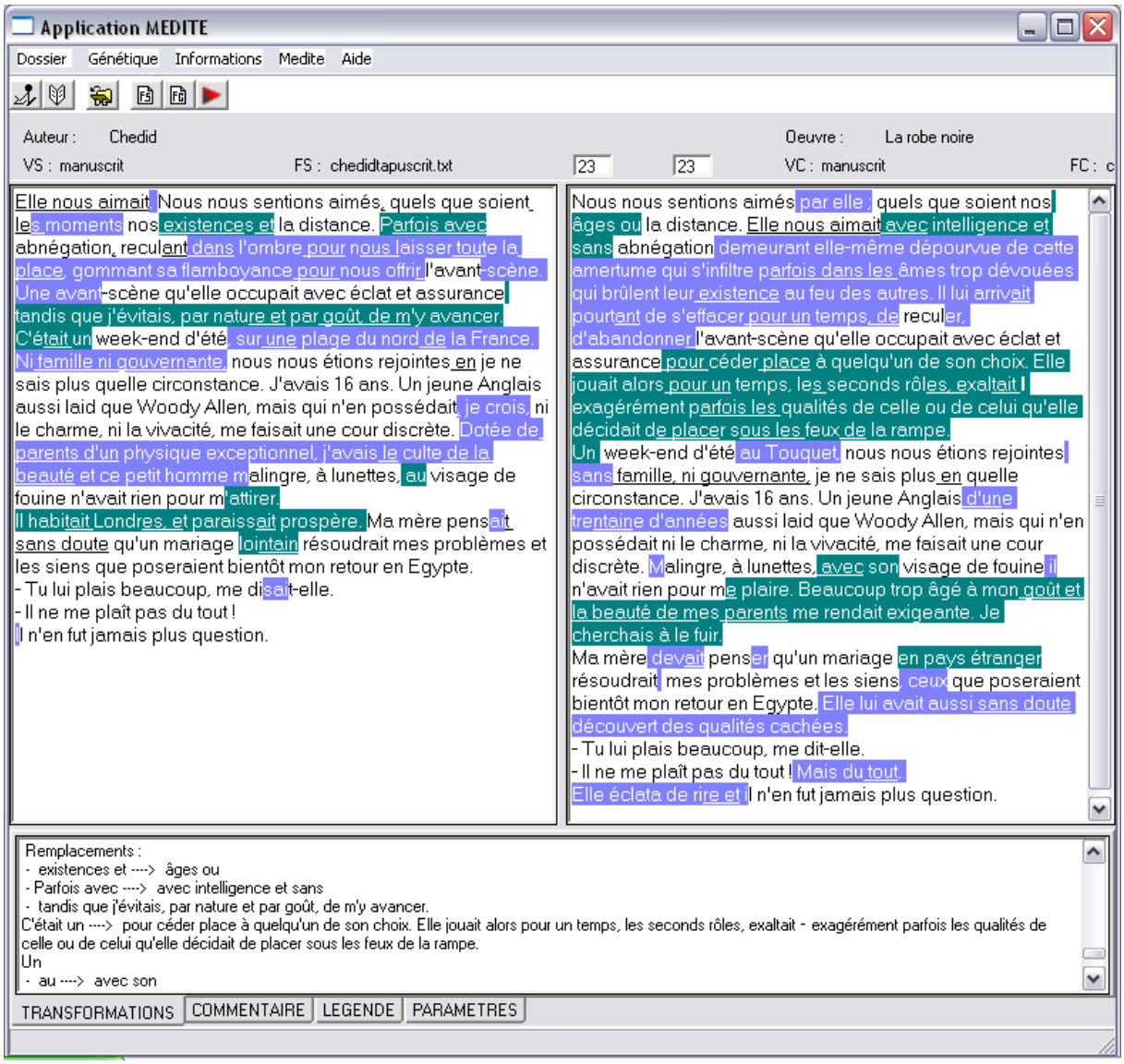


Figure 3. Interface de visualisation de MEDITE

Au terme de cette présentation de l'interface de MEDITE, il faut souligner qu'une fois le travail terminé, l'ensemble des textes, des blocs détectés et des commentaires, sont stockés dans un fichier XML dont le contenu peut être exploité par des procédures d'analyse statistique. De plus, ce fichier peut être rechargé automatiquement dans l'interface, ce qui évite d'avoir à exécuter plusieurs fois l'algorithme de comparaison sur de longs textes comme des romans ou des essais philosophiques.

5. Interprétation génétique

5.1. *Analyse d'un passage : une nouvelle d'Andrée Chedid* La robe noire

Reportons-nous de nouveau à la copie d'écran de la figure trois et observons non plus la forme de l'interface, mais le contenu des deux fenêtres supérieures. La fenêtre gauche contient la tapuscrit du brouillon d'Andrée Chedid donné dans la figure 1 (nous dirons qu'il s'agit là de l'état 1 du texte) et la fenêtre de droite, l'état final de ce brouillon, à l'issue de toutes les campagnes de récriture.

L'application MEDITE permet une visualisation immédiate de tout ce qui se passe entre l'état initial et l'état final du texte de ce brouillon. L'exhaustivité des transformations nous est matériellement donnée dans le tableau inférieur. Voilà, immédiatement offert et classé, le matériau minimum nécessaire au généticien du texte.

Pour ce qui est de ce passage, nous remarquons que les transformations – dont les éléments qui les constituent sont de différentes natures linguistiques – convergent quasiment toutes vers le personnage de la mère « Elle ».

Plus exactement, dans ce passage de l'état initial à l'état final s'opère une vraie conversion dans la façon dont est vue et présentée la place de la mère. Si l'état initial s'efforce de mettre en valeur une générosité maternelle, la suite des suppressions et remplacements montre clairement que l'état final installe une vision toute différente où la mère s'impose et prend toute la place : le bloc « l'avant-scène qu'elle occupait avec éclat et assurance » reste inchangé alors que les segments contextuels, à droite comme à gauche, passent d'une minimisation à une affirmation amplifiante de cette place (« elle » → « par elle » ; « avec abnégation » → « sans abnégation » ; « reculant dans l'ombre pour nous laisser toute la place » → « il lui arrivait pourtant de s'effacer pour un temps » ; insertion de « céder place à quelqu'un de son choix », etc.).

5.2. *Analyse d'un dossier complet. Point de vue global sur la genèse d'un texte : une nouvelle de Pascal Quignard,* Bernon l'enfant.

L'analyse par MEDITE des 5 versions de cette nouvelle dans le déploiement de ses différents états (15 états différents au total inégalement répartis selon les versions) fait apparaître 3 grandes campagnes d'écriture à quoi s'ajoute une mise au net (V5).

– V1 à V2 : campagne que l'on pourrait qualifier de « linguistique » : l'auteur joue sur la perfectibilité de l'écriture, le texte est « amélioré » du point de vue des normes linguistiques.

Ainsi les changements suivants : « du visage » → « de son visage » ; « cette promesse ne tombe pas d'oreille d'un sourd » → « cette promesse ne tombe pas dans l'oreille d'un sourd »

– V2 à V3 : campagne d'écriture focalisée sur une thématique, celle de la beauté. Stylistiquement se développe une isotopie du *beau* (voir figure 4).

Sur la copie d'écran présentée ci-dessus (voir figure 4), le relevé des insertions fait clairement apparaître une véritable série lexicale autour de ce paradigme sémantique : insertions de « qui était très beau », « le Beau Palaiseau », « Beau », « Son bel air s'ajoute à sa beauté », « le bel adolescent », « Une fois pesé, radieux, le bel adolescent »…

Des tournures sont transformées de façon esthétisante : « le crâne » est remplacé par « la tête de mort », « l'église » par « la chantrerie de la basilique », « Il » par « Le Palaiseau », …

Enfin la *visibilité* des transformations textuelles par MEDITE fait "entendre" le travail sur les sonorités.

– V3 à V4 : campagne d'écriture thématique à nouveau autour de la passion. L'examen de la liste des transformations montre plusieurs insertions de modalités à caractère affectif et passionnel (« passionnément catholique », « déteste », « hait »).

L'interprétation de ces thèmes fait apparaître une distinction – sinon une opposition – entre « Catholiques » et « Réformés ».

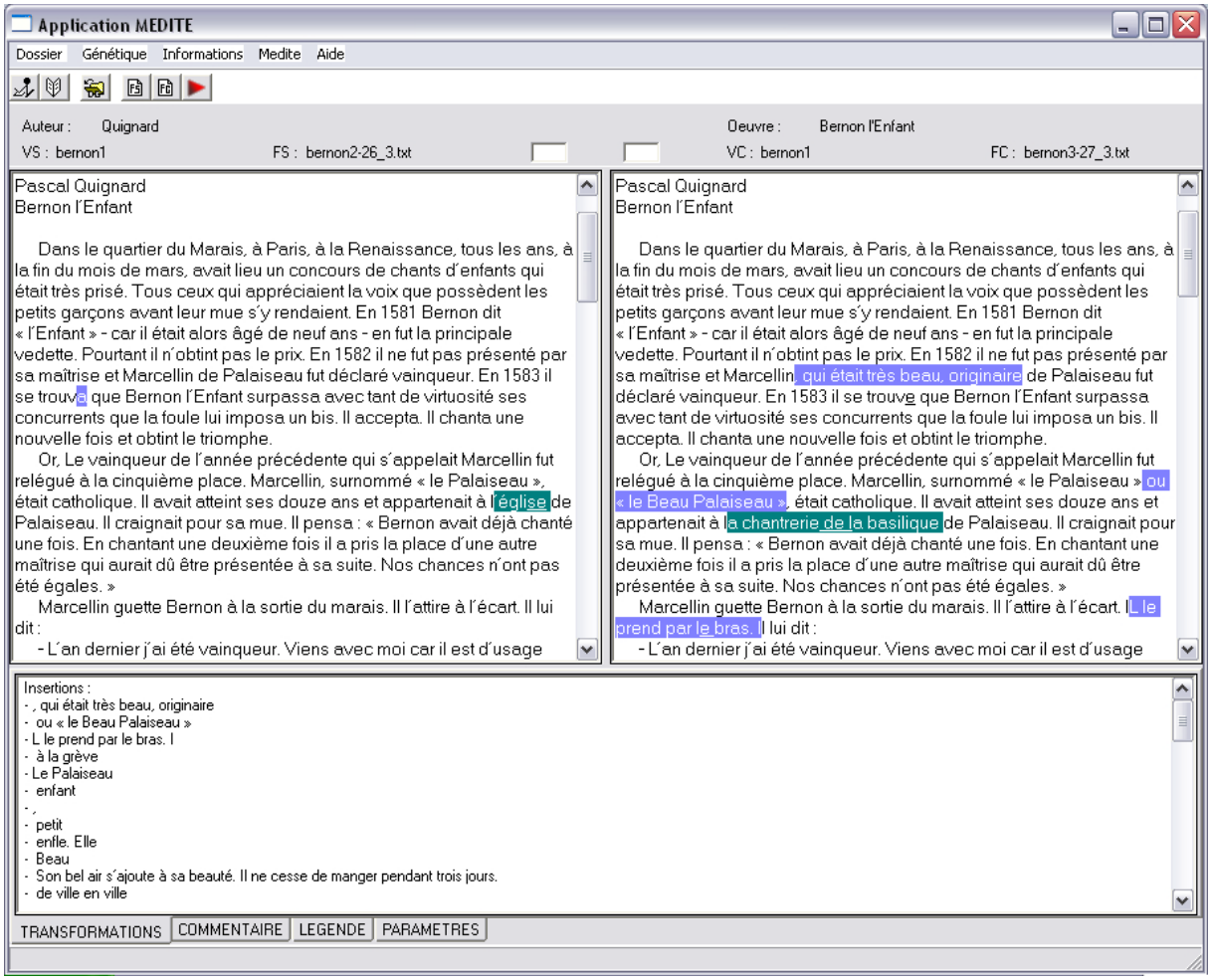


Figure 4. Activation de MEDITE sur les versions 2 et 3 de « Bernon l'enfant »

Cette analyse par MEDITE est tout à fait novatrice pour la génétique des textes. En effet, jusque là, l'étude de la genèse était centrée sur un calibrage qualitatif et plutôt exemplaire, faute de possibilité de relevés exhaustifs des données ; en conséquence, on renonçait soit à l'examen systématique d'un dossier complet dès que celui-ci était trop long, soit à la vision d'ensemble de la genèse. Et, de toute façon, on demeurait dans l'incapacité de fournir des « preuves » fondées sur comptabilisation précise.

Par le biais de MEDITE le généticien dispose d'un matériau *exhaustif*, immédiatement *visible* et surtout dont la comparaison, pièce à pièce, est *directement accessible* et productive.

6. Conclusion

MEDITE est programmé en Python. Il fonctionne actuellement sous les systèmes d'exploitation Windows et LINUX. Une version Mac OS X devrait voir le jour dans les prochains mois. Plusieurs corpus sont actuellement à l'étude avec ce logiciel : Andrée Chedid, *La robe noire*, Louis Althusser *Freud et Lacan*, Marcel Proust *Cahiers*, Pascal Quignard *Bernon l'enfant*. Dès à présent, le logiciel permet d'effectuer automatiquement des études trop fastidieuses pour être réalisées manuellement. Et, même sur des exemples aussi brefs que sur le début de *La robe noire*, ou sur la nouvelle de Pascal Quignard *Bernon l'enfant*, l'observation des transformations explicitées par MEDITE montre à l'évidence la signification du travail de l'auteur ; toutes les réécritures semblent aller dans le même sens : dans le cas du texte

d'Andrée Chedid, la jeune fille est progressivement dessaisie de toutes prérogatives ; au fil des réécritures, elle agit de moins en moins, tandis que la mère, apparemment aimante, la manipule de plus en plus... L'étude de la nouvelle de Pascal Quignard, quant à elle, fait apparaître plusieurs phases distinctes dans le travail de récriture. À ces interprétations sémantiques qui demeurent somme toute assez subjectives, on peut ajouter des études statistiques qui se font sur les transformations elles-mêmes. On devrait donc, grâce au logiciel MEDITE, ouvrir sur une linguistique de l'écrit à même d'aborder quantitativement le travail de réécriture des auteurs. C'est là un premier pas vers de nouvelles applications de l'analyse de données textuelles à la philologie.

Références

de Biasi P.-M. (2000). *La génétique des textes*. Nathan Université.

Cerquiglini B. (1989). *Éloge de la variante. Histoire critique de la philologie*. Seuil.

Chedid A. (1996). La robe noire. In *Les saisons de passage*. Flammarion.

Contat M. et Ferrer D. (Ed.) (1998). *Pourquoi la critique génétique ? Méthodes, théories*, CNRS éditions.

Fenoglio I (2001). Énonciation et genèse dans les autobiographies d'Althusser. *Genesis*, vol. (17) :131-150.

Fenoglio I. et Boucheron S. (Eds) (2002). Processus d'écriture et marques linguistiques. *Langages* vol. (147).

Ganascia J.-G. (2001). Extraction of Recurrent Patterns from Stratified Ordered Trees. In *Actes de la conférence ECML*. Springer.

Grésillon A. (1994). *Éléments de critique génétique*. PUF.

Hay L. (2002). *La littérature des écrivains. Études de critique génétique*. Corti.

Crochemore M. et Rytter W. (1994). Text Algorithms. *Approximate pattern matching* : 237-251.

Karp R.M., Miller R.E. et Rosenberg A.L. (1972). *Rapid Identification of Repeated Patterns in Strings, Trees and Arrays*. In *Proceedings of the 4th Annu. ACM Symp. Theory of Computing* : 125-136.

Landraud A.-M., Avril J.-F. et Chrétienne P. (1989). An algorithm for Finding a Common Structure Shared by a Family of Strings. *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. (11 : 8) : 890-895.

Lebrave J.-L. (1984). Le traitement automatique des brouillons. *Programmation et sciences humaines* (N° spécial), MSH.

Lebrave J.-L. (1990). *Déchiffrer, transcrire, éditer la genèse*. *Proust à la lettre.* Du Lérot : 141-162.

Sankoff D. et Kruskal J.B. (1983). *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, Reading.

Analyse exploratoire d'entrevues de groupe : les jeunes Français et le tabac

Claire Gélinas-Chebat[1], François Daoust[2], Monique Dufresne[3], Karine Gallopel[4], Marie Élaine Lebel[5]

[1]Professeure, DLDL, UQAM – Montréal – Canada – chebat.claire@uqam.ca
[2]Informaticien au Centre ATO, doctorant U. de Franche-Comté – Besançon – France
daoust.francois@uqam.ca
[3]Professeure associée, DLDL, UQAM – Montréal – Canada
[4]Maître de conférences, Université de Rennes – Rennes – France
[5]Chercheure, DLDL, UQAM – Montréal – Canada

Abstract

This paper presents a methodology related to text analysis constituted from a teenager group of discussion about tobacco usage. The emphasis of the methodology point out a construct of a content analysis based on iteractive process, which leads to data base building and hypothesis-testing statistical analysis. Statistical comparison in between lexical data from sub-categorization of the text associated with sociological information permit us to lead and relate our research questions to the textual analysis.

Résumé

Cet article a pour objectif de montrer, au moyen d'une analyse exploratoire d'un corpus constitué de transcriptions d'entrevues de groupe sur l'usage du tabac, une méthodologie d'analyse textuelle qui vise à construire un système de catégories de manière itérative. Des algorithmes simples de comparaison statistique entre lexiques associés à des sous-textes correspondant à nos variables sociologiques permettent d'appuyer la construction d'un système catégoriel qui établit un pont entre la problématique de recherche et nos données textuelles.

Mots-clés : analyse de textes, corpus d'entrevues de groupes, catégorisation, approche inductive et itérative, analyse lexicale.

1. Introduction

Nous présentons ici une analyse d'entrevues de groupe sur l'usage et les attitudes à l'égard de la consommation du tabac. Il s'agit d'un traitement exploratoire de transcriptions en vue de mettre en place, dans un premier temps, une méthodologie permettant le traitement textuel et itératif des données et, dans un deuxième temps, de s'assurer d'un protocole expérimental adapté à la saisie de nouvelles données dans un contexte similaire. En effet, cette recherche constitue une étape préparatoire à la saisie complémentaire de données du même type, cette recherche est le premier volet d'un projet de recherche plus vaste qui se réalisera dans un contexte français, canadien et américain.

Outre le traitement lexical, le traitement exploratoire de nos données tient compte du profil sociologique de nos sujets afin de proposer une grille catégorielle qui permet de comprendre, à travers les interventions de chacun, l'influence de messages d'avertissement sur l'usage du tabac. Les résultats de cette analyse seront exploités plus à fond dans les autres volets de cette recherche.

La première partie décrit notre problématique et aborde la question de l'attitude des adolescents à l'égard du tabagisme et des messages dissuasifs. Nous enchaînons sur les caractéristiques de notre corpus à la section suivante. La quatrième partie expose notre méthodologie interactive d'analyse textuelle. Enfin, la dernière partie présente des extraits du tableau de bord de la recherche montrant à l'œuvre notre démarche dans ses phases inductive et hypothético-déductive.

2. Problématique et questions de recherche

Cette recherche tente de comprendre la portée des messages antitabac chez les adolescents avec l'objectif à long terme de réduire significativement leur consommation de cigarettes. Malgré des efforts marqués pour dissuader les adultes à fumer et convaincre les jeunes à ne pas commencer à fumer, la consommation des produits du tabac reste très importante (Santé Canada *et al.*, 1999). En France, près d'un jeune sur deux fume.

Les adolescents minimisent les risques de l'usage des produits dangereux et tendent à sous-estimer les dangers de l'usage du tabac (Leventhal *et al.*, 1987). Quel discours doit-on tenir dans ces messages d'avertissement si on veut produire des messages efficaces ? Une étude comparative de différentes recherches empiriques sur les effets des menaces dans le domaine des mises en garde sur la santé montre que plus le message suscite des émotions de peur, plus les effets sur l'attitude, l'intention et les changements de comportements sont grands. De même, plus la sévérité du message est forte, plus l'attitude, l'intention et les comportements changent (Witte et Allen, 2000). Mais le goût du risque, par exemple des sports extrêmes et l'attrait du « fruit défendu » (Parker-Pope, 1997) n'a-t-il pas l'effet contraire à l'effet désiré ?

Dans la démarche exploratoire, nous avons traité les données recueillies dans le cadre d'une discussion ; des groupes de plusieurs participants discutaient sur le thème de la cigarette, puis ils étaient exposés à différents messages d'avertissement avec des menaces graduées. Chaque groupe était exposé à un seul message d'avertissement. À cette étape-ci de la recherche, nous nous sommes mis à l'écoute des jeunes exposés à ces différents messages dans le but de reprendre les arguments évoqués ultérieurement afin de construire dans une deuxième phase expérimentale des messages persuasifs dont nous mesurerons expérimentalement les effets.

3. Un corpus de groupes de discussion

Ce corpus comprend neuf entrevues sur le tabagisme chez les jeunes et leur perception de la publicité antitabac. Elles ont été réalisées à Rennes en 2000 auprès de 48 jeunes Français qui, pour la plupart, fréquentent une institution scolaire et qui sont âgés de 15 à 25 ans. Chacune des séances réunit 5-6 jeunes (fumeurs ou non-fumeurs, hommes et femmes) et un intervenant, et se divise en deux parties. La première partie se déroule après que l'intervenant a posé quelques questions pour amorcer la discussion et la seconde se caractérise par l'introduction d'une brochure. Il existe différentes versions de la brochure selon deux paramètres : les effets du tabagisme sur la santé et les solutions pour arrêter de fumer. Ces deux paramètres se présentent comme suit, trois niveaux de menace (faible, moyen et fort), et deux niveaux de solution (faible et fort).

Les entrevues ont été enregistrées sur bandes audio et retranscrites en format Word. Au début de chaque transcription, les données sociologiques des personnes qui participent à l'entrevue sont précisées : âge, sexe, fumeur/non-fumeur. Nous avons effectué l'analyse du corpus au moyen du logiciel SATO (Daoust, 1996). Les annotations éditiques ont été remplacées par un balisage symbolique conforme à la syntaxe de SATO. Un astérisque introduit les balises,

aussi appelées *propriétés*, et celles de notre corpus sont les suivantes : *locuteur, *sexe, *fumeur, *page et *thème. En voici un exemple :

***page=gallo02/11**

***thème=brochure (…) *locuteur=s36 *fumeur=non *sexe=ho** Bah, la brochure là, elle nous présente ce qui nous attend si on fume. Mais c'est très... quoi, moi j'ai lu ça, mais je ne sais pas je ne suis pas fumeur, donc je ne ressens peut-être pas ça de la même façon. À la limite on passe dessus comme ça, ça apporte quelques chiffres.

4. Une construction itérative des catégories

Nous avons utilisé une démarche exploratoire inspirée de l'approche adoptée pour le corpus *Message d'amour* (Daoust, 1999) et qui vise à *faire parler les données*. Le principe de base de la démarche consiste à comparer, avec des indices statistiques simples, les lexiques associés à des sous-textes découpés d'après nos variables sociologiques. Cette comparaison sera reprise de façon itérative de façon à s'appuyer sur le lexique brut pour construire un lexique catégorisé reflétant les points d'ancrage de notre chaîne interprétative.

Le découpage du texte et la constitution des lexiques associés procèdent comme suit. Les balises (*propriétés*) introduites dans notre corpus permettent de segmenter le corpus en opposant, par exemple, l'ensemble des interventions *avant* et *après* la présentation du document publicitaire. De la même façon, on peut découper le corpus entre, d'une part, les interventions des hommes et, d'autre part, celles des femmes, excluant les interventions des modérateurs. Ces balises étant indépendantes, elles peuvent être combinées à loisir lors de l'exploration du corpus. Ainsi, on pourrait comparer les interventions des hommes seulement, ou des femmes seulement, *avant* et *après* la présentation de la brochure pour voir si la réaction à la brochure dépend du sexe des sujets.

Pour comparer nos lexiques, nous utilisons un *algorithme de distance lexicale* basé sur la *distance du Chi2*. La mesure évalue l'écart dans l'utilisation d'un vocabulaire donné entre deux sous-ensembles du corpus. Les formes lexicales sont triées par ordre décroissant de contribution à la mesure de distance, ce qui permet d'identifier, par ordre d'importance, les spécificités de chaque sous-texte. L'algorithme peut être appliqué aux formes lexicales elles-mêmes ou aux valeurs de propriétés correspondant à notre catégorisation lexicale. Ici, l'approche est essentiellement dichotomique : on compare un sous-texte à un autre, via leur lexique respectif. On peut aussi avoir recours à un *algorithme de participation* qui calculera les moyennes normalisées d'un ensemble de formes lexicales, correspondant généralement à une catégorie lexicale, pour chacun des sous-textes constitués en cours d'analyse.

Notre démarche exploratoire est fondée sur un va et vient interactif entre ce que nous révèle l'analyse lexicale et les contextes d'utilisation des mots mis en évidence par les algorithmes de distance et de participation. Dans ce premier temps de l'analyse, nous privilégions une approche univariée afin de mieux saisir la spécificité de la stratification induite par chacune de nos variables sociologiques. Dans un deuxième temps, on peut, avec les mêmes outils, comparer des sous-textes faisant appel à plus d'une variable, comme dans l'exemple ci-haut tenant compte du sexe et de l'introduction de la brochure.

Au premier niveau de l'analyse, nous travaillons sur les données brutes, c'est-à-dire les formes lexicales elles-mêmes. On se donne ainsi la possibilité de voir apparaître des différenciations portées par la morphologie des mots en termes de nombre, genre, personne, temps. On s'intéresse tout autant, sinon davantage, à des marqueurs d'énonciation, comme les pronoms personnels, les marqueurs phatiques, les marques de la négation, de l'interrogation, les verbes

épistémiques (croire, penser...), etc. qu'aux termes pleins. C'est en s'appuyant sur l'analyse lexicale des données brutes que nous élaborerons nos grilles catégorielles.

Le retour constant aux énoncés, ne serait-ce que par un parcours rapide des contextes courts de type KWIC (*key words in context*), est cependant essentiel pour ébaucher nos hypothèses sur le fonctionnement du discours et le positionnement des locuteurs d'après leur profil social. Ce va et vient entre l'analyse lexicale et les énoncés permet en effet d'inscrire les unités lexicales dans des systèmes de catégories sémantiques et énonciatives susceptibles de traduire, dans le discours même, ce que l'on cherche à comprendre, à savoir ici l'attitude des jeunes par rapport au tabagisme et l'influence de publicités dissuasives. La catégorisation vise donc à établir le pont entre la problématique de recherche et nos données textuelles. Elle correspond un peu à la procédure de codage de l'analyse qualitative à cette différence qu'elle s'appuie sur des procédures d'analyse lexicale qui permettent de tenir compte de l'ensemble des données et sur l'examen de phénomènes discursifs difficilement repérables par une simple lecture linéaire.

La reprise des analyses univariées et multivariées sur les données catégorisées s'inscrit dans les procédures de validation de nos hypothèses interprétatives. Notre première approche, inductive, qui implique à la fois une sensibilité aux procédés linguistiques et à la problématique de la recherche, sera donc relayée par une approche davantage hypothético-déductive.

5. Tableaux... d'une exploration

L'analyse de distance permet de déterminer le vocabulaire qui caractérise un sous-texte, c'est-à-dire les formes qui marquent davantage l'originalité du vocabulaire d'une partie du corpus par rapport à l'autre. Dans le tableau qui suit (tableau I), les mots qui caractérisent le plus le discours avant l'introduction de la brochure sont suivis d'un astérisque. Les mots sans astérisque caractérisent davantage les propos tenus après l'introduction de la brochure.

	*				
Fréqtot	A	B	explique	cumul	
0.07	0.14	0.02	0.44	0.44	clair *
0.23	0.38	0.18	0.40	0.84	aussi *
0.05	0.11	0.02	0.31	1.15	plaisir *
0.06	0.11	0.02	0.31	1.46	dépendance *
0.02	0.00	0.05	0.28	1.75	témoignage
0.09	0.04	0.15	0.28	2.02	"
0.01	0.03	0.00	0.26	2.28	3ème *
0.02	0.05	0.00	0.25	2.54	doigts *
0.06	0.01	0.09	0.24	2.78	risques
0.02	0.05	0.00	0.24	3.02	primaire *
0.37	0.45	0.25	0.24	3.25	ils *
0.59	0.62	0.87	0.23	3.49	j'
0.03	0.01	0.06	0.23	3.72	concret
0.01	0.00	0.04	0.23	3.95	cinq
0.09	0.13	0.04	0.22	4.17	santé *
0.02	0.00	0.04	0.21	4.38	solution
0.02	0.04	0.00	0.20	4.58	appelle *
0.02	0.00	0.05	0.20	4.78	chiffres
0.03	0.01	0.06	0.20	4.98	routière

0.01	0.03	0.00	0.19	5.17	choqué *
0.01	0.03	0.00	0.19	5.37	influencé *
0.01	0.03	0.00	0.19	5.56	dents *
0.15	0.10	0.21	0.19	5.74	elle
0.36	0.36	0.53	0.18	5.93	!
0.01	0.00	0.03	0.18	6.11	morts
0.03	0.02	0.06	0.18	6.29	y'
0.02	0.04	0.00	0.18	6.47	dérange *
0.07	0.04	0.11	0.18	6.65	Cela
0.28	0.17	0.32	0.18	6.83	Te
0.02	0.00	0.04	0.18	7.00	Image
0.02	0.00	0.04	0.18	7.18	Provoque
0.02	0.04	0.00	0.17	7.35	odeur *
0.04	0.09	0.03	0.17	7.52	effectivement *
0.02	0.04	0.00	0.17	7.69	jaunes *
0.16	0.17	0.06	0.17	7.86	toi *
0.06	0.10	0.03	0.17	8.03	niveau *
0.01	0.03	0.00	0.17	8.20	publicitaire *
0.02	0.00	0.04	0.17	8.36	Long

Tableau 1. Analyse de distance sur les formes lexicales brutes avant/après l'introduction de la brochure

Si on s'attarde aux items lexicaux pleins, c'est-à-dire aux noms, adjectifs et aux verbes, il semble que les mots qui décrivent l'apparence physique et la santé en général sont ceux qui caractérisent le plus vocabulaire avant l'introduction de la brochure *clair, doigts, dents, santé* sans oublier la notion de *plaisir* et de *dépendance ;* après l'introduction de la brochure, il est remarquable de constater que les mots *témoignage, concret, solution, chiffres, mort* sont ceux qui apparaissent en tête de liste. Il appert que les deux sous textes ne font pas ressortir les effets du tabac dans les mêmes termes. Avant on parle de plaisir et des effets néfastes sur la santé et particulièrement sur l'apparence physique, les dents et les doigts jaunes et sur la dépendance. Après, les effets font toujours partie du discours, mais alors non plus en termes de plaisir, mais en termes de risque et de mort. Notons par ailleurs la présence importante du pronom *j'* après l'introduction de la brochure. Ceci pourrait suggérer que la brochure provoque une plus grande implication personnelle de nos sujets.

Pour valider ces observations, la première stratégie de vérification consiste à parcourir rapidement les contextes. Ainsi, nous avons constaté que le mot *clair* n'a rien à voir avec l'apparence, mais est plutôt utilisé comme marque évaluative : *C'est clair, c'est évident*.

Nous avons poursuivi nos travaux en nous donnant une procédure de vérification des hypothèses d'interprétation construites à partir des premiers résultats obtenus lors de l'analyse de distance sur les unités lexicales brutes. C'est ici qu'entre en jeu la catégorisation des unités lexicales. Il s'agit de déterminer avec plus de raffinement les sujets traités au cours des discussions, toujours dans la perspective de déterminer si l'introduction de la brochure produit des changements dans le discours des participants.

Nous avons donc introduit une propriété lexicale que nous avons nommée *sujet*. Cette propriété englobe en fait plusieurs grilles d'analyse que nous aurions pu regrouper dans des propriétés différentes. Mais dans l'idée de procéder de façon itérative par raffinements successifs de nos procédures, il était suffisant, à ce stade-ci, de n'avoir qu'un système de catégories dont certaines peuvent déjà faire l'objet d'une structuration plus fine. Ainsi, seront rajoutées à

notre grille cinq catégories (*Soc-X*) associées aux formes lexicales décrivant l'environnement social. Pourquoi ? Nous avions remarqué au cours de l'analyse de distance que les items lexicaux qui faisaient référence à l'aspect social du tabagisme, c'est-à-dire de ses conséquences sur les rapports sociaux des jeunes, ressortaient beaucoup. On a donc établi en suffixe une échelle décrivant le niveau de l'environnement social, allant du plus intime au plus général : *soc-je, soc-ami, soc-famille, soc-jeune, soc-gens*.

Voici le descriptif de la propriété *sujet*. On a déterminé *28 sujets : apparence, arrêt, négation, concret, danger, dépendance, soc-je, maladie, mort, plaisir, publicité, tabac, nicotine, drogue, interdiction, fumeur, soc-ami, soc-famille, soc-gens, liberté, envie, conscience, volonté, soc-jeune, coûts, début, santé, éducation, prévention*.

La procédure de catégorisation procède des mots caractéristiques révélés par l'algorithme de distance, vers l'ensemble du vocabulaire. Après ces mots, nous avons examiné de façon systématique les mots fréquents et complété la catégorisation en examinant le lexique trié par ordre alphabétique pour catégoriser les variantes flexionnelles pertinentes. Pour confirmer nos intuitions, nous reprenons les analyses statistiques sur les valeurs de la propriété *sujet* comme l'illustre le tableau II qui suit.

	*				
Fréqtot	A	B	explique	cumul	
0.21	0.43	0.11	31.23	31.23	apparence *
0.09	0.02	0.16	13.85	45.08	Concret
0.08	0.14	0.05	6.75	51.83	plaisir *
0.13	0.21	0.10	6.63	58.46	dépendance *
0.14	0.19	0.08	5.64	64.10	santé *
0.11	0.17	0.08	5.39	69.49	éducation *
0.18	0.11	0.22	5.12	74.61	Volonté
0.10	0.08	0.17	4.75	79.36	Mort
1.95	2.19	1.82	4.53	83.89	tabac *
0.05	0.10	0.05	3.26	87.15	soc-ami *
0.17	0.25	0.16	3.12	90.27	coûts *
0.32	0.28	0.40	2.81	93.09	Maladie
0.75	0.59	0.72	1.44	94.53	Publicité
0.21	0.26	0.20	1.37	95.90	soc-famille *
0.11	0.14	0.11	0.84	96.74	drogue *
0.20	0.22	0.17	0.82	97.55	liberté *
0.74	0.69	0.78	0.67	98.23	soc-gens
0.17	0.14	0.18	0.66	98.89	Envie
0.31	0.29	0.24	0.54	99.44	soc-jeune *
0.05	0.08	0.06	0.19	99.62	nicotine *
0.63	0.60	0.64	0.18	99.81	Arrêt
0.08	0.07	0.08	0.09	99.90	Conscience
0.24	0.17	0.19	0.03	99.93	Danger
0.22	0.17	0.16	0.03	99.96	début *
0.13	0.11	0.12	0.02	99.98	Prévention
2.28	2.48	2.50	0.01	99.99	Négation
0.48	0.44	0.45	0.01	100.00	Fumeur
2.14	2.54	2.53	0.00	100.00	soc-je *

Tableau 2. Analyse de distance sur les formes de la catégorie sujet *avant/après l'introduction de la brochure*

La notion d'*apparence* se confirme. Les sujets, avant la brochure, abordent les effets superficiels du tabagisme, à savoir, la couleur des dents et des doigts, le teint, l'odeur des vêtements et des cheveux… La notion de plaisir ressort aussi comme thème avant l'introduction de la brochure, ainsi que les notions de dépendance, santé et éducation.

Après l'introduction de la brochure, la catégorie *concret* (impact et solutions) ressort. Nous voyons aussi émerger les notions de *volonté, mort* et *maladie*. D'autre part, nous constatons que notre hypothèse sur les pronoms personnels de la première personne (*soc-je*) ne se confirme pas. L'écart observé avant et après la brochure était spécifique à la forme *j'*.

Nous affinons l'analyse en comparant les interventions avant et après l'introduction de la brochure selon le profil sociologique des sujets. Ainsi, en comparant les données qui suivent (tableau III), il est remarquable de constater que ce sont les non-fumeurs qui semblent le plus touchés par la brochure comme en témoigne la dominance des thèmes relatifs aux effets négatifs du tabagisme : *maladie* et *mort*.

Comparaison « avant – après » pour les fumeurs						**Comparaison « avant – après » pour les non-fumeurs**					
Mode propriété sujet						**Mode propriété sujet**					
Fréqtot	Afu	Bfu	explique	cumul		Fréqtot	Anf	Bnf	explique	cumul	
0.21	0.47	0.11	32.55	32.55	apparence*	0.75	0.42	0.99	15.22	15.22	publicité
0.09	0.03	0.19	18.05	50.60	concret	0.21	0.38	0.10	12.52	27.74	apparence*
0.11	0.20	0.06	10.15	60.75	éducation*	0.05	0.14	0.02	10.64	38.38	soc-ami*
0.13	0.20	0.08	6.26	67.01	dépendance*	0.17	0.37	0.15	9.73	48.11	coûts*
0.18	0.07	0.21	5.79	72.79	volonté	0.14	0.20	0.03	8.21	56.32	santé*
0.08	0.15	0.06	4.85	77.64	plaisir*	0.32	0.29	0.54	6.94	63.26	maladie
0.48	0.35	0.53	3.96	81.60	fumeur	0.10	0.06	0.19	5.96	69.22	mort
0.17	0.15	0.25	3.77	85.37	envie	0.08	0.14	0.04	4.51	73.72	plaisir*
0.75	0.71	0.51	2.70	88.07	publicité*	1.95	2.32	1.82	4.38	78.11	tabac*
1.95	2.09	1.81	2.23	90.30	tabac*	0.09	0.01	0.11	4.01	82.11	concret
0.21	0.28	0.20	1.79	92.08	soc-famille*	0.48	0.57	0.34	4.00	86.12	fumeur*
0.10	0.09	0.15	1.74	93.83	mort	0.13	0.24	0.13	3.15	89.27	dépendance*
0.14	0.18	0.12	1.60	95.43	santé*	0.11	0.19	0.11	2.32	91.59	drogue*
0.22	0.18	0.12	0.99	96.42	début*	0.20	0.20	0.13	1.09	92.68	liberté*
0.63	0.57	0.67	0.98	97.41	arrêt	0.05	0.08	0.04	1.05	93.73	nicotine *
2.14	2.74	2.93	0.91	98.32	soc-je	2.14	2.25	2.01	0.93	94.66	soc-je *

Tableau 3. Analyse de distance avant/après pour les fumeurs et les non-fumeurs

Une première analyse des résultats obtenus nous a amenés à conclure que les hommes semblaient plus interpellés par l'introduction de la brochure que les femmes. Ce résultat confirme-t-il les autres analyses sur les effets des campagnes antitabac, à savoir que les femmes sont moins touchées que les hommes ? Une réponse positive, à cette étape de l'analyse, serait prématurée.

Une autre façon de visualiser les résultats nous est donnée par l'analyseur PARTICIPATION. Pour une catégorie donnée, l'analyseur calcule sa fréquence relative dans les divers sous-tex-

tes, ce qui permet de voir si les variables sociologiques ont une influence sur l'utilisation des mots catégorisés. Les tableaux IV et V illustrent la distribution des catégories *apparence* et *mort* dans le corpus complet et divers sous-textes. **A** et **B** désignent *avant* et *après* la brochure. Nous avons aussi les particules **fu** et **nf** pour *fumeur* et *non-fumeur*, ainsi que **ho** et **fe** pour *homme* et *femme*.

Propriété	Couverture	Lexèmes	Occurrences	Cote Z
Fréqtot	78703/78703 (100.00%)	37/3985 (0.93%)	168/78703 (0.21%)	0.00
A	23544/78703 (29.91%)	30/2087 (1.44%)	101/23544 (0.43%)	7.17
B	28074/78703 (35.67%)	18/2351 (0.77%)	30/28074 (0.11%)	-3.87
Afu	13758/78703 (17.48%)	24/1580 (1.52%)	64/13758 (0.47%)	6.40
Bfu	15923/78703 (20.23%)	13/1749 (0.74%)	18/15923 (0.11%)	-2.75
Anf	9786/78703 (12.43%)	19/1240 (1.53%)	37/9786 (0.38%)	3.53
Bnf	11898/78703 (15.12%)	8/1425 (0.56%)	12/11898 (0.10%)	-2.66
Aho	14468/78703 (18.38%)	16/163 (4 0.98%)	44/14468 (0.30%)	2.36
Bho	16010/78703 (20.34%)	11/1797 (0.61%)	19/16010 (0.12%)	-2.60
Afe	9076/78703 (11.53%)	24/1153 (2.08%)	57/9076 (0.63%)	8.56
Bfe	11811/78703 (15.01%)	9/1379 (0.65%)	11/11811 (0.09%)	-2.83

Tableau 4. Analyseur PARTICIPATION (sujet=apparence)

Propriété	Couverture	Lexèmes	Occurrences	Cote Z
Fréqtot	78703/78703 (100.00%)	9/3985 (0.23%)	80/78703 (0.10%)	0.00
A	23544/78703 (29.91%)	4/2087 (0.19%)	19/235440 (0.8%)	-1.01
B	28074/7870335 (67%)	6/2351 (0.26%)	47/28074 (0.17%)	3.46
Afu	13758/78703 (17.48%)	4/1580 (0.25%)	13/13758 (0.09%)	-0.26
Bfu	15923/78703 (20.23%)	6/17490.(34%)	24/15923 (0.15%)	1.94
Anf	9786/7870312.(43%)	2/1240 (0.16%)	6/9786 (0.06%)	-1.25
Bnf	11898/78703 (15.12%)	3/1425 (0.21%)	23/11898 (0.19%)	3.14
Aho	14468/78703 (18.38%)	4/1634 (0.24%)	8/14468 (0.06%)	-1.75
Bho	16010/78703 (20.34%)	4/1797 (0.22 %)	21/16010 (0.13%)	1.17
Afe	9076/78703 (11.53%)	2/1153 (0.17%)	11/9076 (0.12%)	0.58
Bfe	11811/78703 (15.01%)	5/1379 (0.36%)	26/1181 (0.22%)	4.04

Tableau 5. Analyseur PARTICIPATION (sujet=mort)

On a repris cette procédure de comparaison des interventions avant et après l'introduction de la brochure, mais en distinguant selon le niveau de *menaces* contenu dans la brochure. On a noté que la présence d'une menace forte provoque un débat sur la liberté et la contrainte. Si on tient compte de la variable *fumeur*, cette préoccupation en vient à occuper le premier rang. Il y a là sans doute des leçons à tirer pour les campagnes antitabac.

Ces quelques exemples, extraits d'une démarche qui sera reprise et développée sur un corpus québécois en cours de constitution, visent d'abord à montrer une approche permettant de construire un protocole de constitution et d'analyse de corpus qui soit à la fois transparent et respectueux de la spécificité du contexte d'énonciation, ici des échanges oraux analysés sous forme de transcriptions. Mentionnons notamment que cette analyse exploratoire a mené à certaines modifications du protocole d'entrevue. Notamment, nous voulions limiter l'introduction par l'intervenant de sujets ou de thèmes qui n'auraient pas été abordés par les jeunes eux-mêmes. Notre protocole expérimental de la deuxième phase de la recherche utilisait des formulations beaucoup moins suggestives qu'au moment de la première phase d'expérimentation. Par exemple, la question *Y a-t-il des conséquences négatives à fumer ?* a été reformulée par *Que pensez-vous que ça apporte de fumer la cigarette ?* Notre première phase de recherche nous a permis de mettre en place une procédure d'analyse textuelle que nous comptons appliquer à nouveau.

Pour conclure, la démarche d'analyse lexicale de nos données est une démarche itérative qui combine l'approche inductive, souvent associée aux méthodes qualitatives, l'utilisation d'outils simples de statistique lexicale, et une approche plus sensible à la pragmatique textuelle. Ce traitement textuel des données a aussi l'avantage de produire des données qualifiées pouvant être soumises à des algorithmes statistiques multivariés susceptibles de mesurer les corrélations entre les variables sociologiques et les différentes modalités catégorielles construites au cours du processus d'analyse textuelle.

Références

Daoust Fr. (1996). *SATO 4, Manuel de référence*. Centre ATO, UQAM, Montréal.

Daoust Fr. (1999). *Corpus Message d'amour : analyse exploratoire*. Centre ATO, Montréal, http://www.ling.uqam.ca/sato/analyses/amour1.html.

Gilmore J. (2000). *Report on Smoking Prevalence in Canada, 1985 to 1999, Statistiques Canada*. Catalogue 82F00077XIE, Ottawa : Ministère de l'Industrie et du commerce du Canada.

Parker-Pope T. (1997). Danger: Warning Labels May Backfire. *Wall Street Journal* (April 28), B1, B8.

Santé Canada, Statistics Canada, Canadian Institute for Health Information (1999). *Statistical Report on the Health of Canadians*. Ottawa. Minister of Public Works and Government Services Canada.

Snyder L.B. et Blood D.J. (1992). Caution: Alcohol and The Surgeon General's Alcohol Warnings May Have Adverse Effects on Young Adults. *Journal of Applied Communication Research*, vol. (20/1): 37-53.

Witte K. et Allen M. (2000). A Meta-Analysis of Fear Appeals: Implications for Effective Public Health Campaigns. *Health Education & Behavior*, vol. (27/5) : 591-616.

D'un dictionnaire de lemmatisation (*D.A.G.*) à un dictionnaire dérivationnel du grec ancien (*D.D.G.*)

Raphaël Gérard, Bastien Kindt

Université catholique de Louvain – Institut orientaliste – Place Blaise Pascal, 1
1348 Louvain-la-Neuve – Belgique
gerard@ori.ucl.ac.be, kindt@ori.ucl.ac.be

Abstract

The lemmas of the *Dictionnaire Automatique Grec* (*D.A.G.*), developed at the U.C.L. for the lemmatisation of patristic and Byzantine Greek texts, are subject of tagging of the constituting morphemes. These data are gathered in a *Dictionnaire Dérivationnel Grec* (*D.D.G.*). By means of specific interfaces, lexical tools make a selection of lemmas with common morphemes possible; a listing of all words – composed or derived – belonging to the same theme may also be visualised as a morphological tree.

Résumé

Les lemmes du *Dictionnaire Automatique Grec* (*D.A.G.*), développé à l'U.C.L. pour la lemmatisation des sources grecques patristiques et byzantines, font l'objet d'un étiquetage systématique des morphèmes qui les constituent. Ces données sont rassemblées dans le *Dictionnaire Dérivationnel Grec* (*D.D.G.*). Une interface d'interrogation permet de sélectionner les lemmes partageant des morphèmes communs. Une autre permet d'afficher sous la forme d'un arbre dérivationnel tous les mots, composés ou dérivés, issus d'un même thème.

Mots-clés : dictionnaire électronique, dictionnaire dérivationnel, étiquetage, grec ancien, lemme, lexique, morphème, récursivité

1. Introduction

1.1. *Le Dictionnaire Automatique Grec (D.A.G.) : présentation*

Le *D.A.G.* est un dictionnaire électronique de la langue grecque ancienne élaboré par deux équipes de l'U.C.L., celle du « Projet de recherche en lexicologie grecque » de l'Institut orientaliste[1], et celle du Centre de Traitement Automatique du Langage (CENTAL)[2]. Utilisé comme dictionnaire de référence pour la lemmatisation automatisée des sources grecques patristiques et byzantines[3], sa nomenclature, structurée en une base de données relationnelle, rassemble les matériaux lexicaux dénombrés au fil des traitements successifs. En tenant compte des analyses passées et en cours, l'ensemble porte sur un *corpus* totalisant 4.284.380 mots-occurrences ; 174.758 formes différentes constituent la microstructure du dictionnaire,

[1] La bibliographie complète du projet est disponible sur la toile à l'adresse http://tpg.fltr.ucl.ac.be ; cf. aussi trois contributions récentes : Coulie (2003), Kindt (2003a) et Kindt (2003b).

[2] L'ancien Centre de Traitement Électronique des Documents (CETEDOC), cf. http://cental.fltr.ucl.ac.be.

[3] Ce choix thématique découle des domaines de recherches propres à l'Institut orientaliste. Les concordances sont publiées dans le *Thesaurus Patrum Graecorum* (*T.P.G.*), une sous-collection du *Corpus Christianorum* diffusée par Brepols Publisher, cf. http://www.brepols.net et http://www.corpuschristianorum.org, ainsi que le site du *T.P.G.* cité note i.

33.874 lemmes, sa macrostructure (cf. fig. 3). Les analyses effectuées portent majoritairement, mais non exclusivement, sur des sources patristiques du IV^e^ s. ap. J.-C. La nature fortement classicisante de la langue de ces textes, ainsi que des principes explicites de formulation des lemmes et la stabilité des normes de dépouillement des formes, rendent l'utilisation du dictionnaire opérationnelle sur des sources antérieures, d'époque classique en l'occurrence, ou postérieures[4]. Depuis 1987[5], il a subi d'importantes modifications susceptibles de lui conférer, progressivement, les potentialités caractéristiques des outils contemporains d'analyse lexicale.

Dans le cadre de la lemmatisation, la structure du *D.A.G.* a été rendue comparable à celle des dictionnaires du Laboratoire d'Automatique Documentaire et Linguistique (L.A.D.L.) de l'Université de Marne-la-Vallée (Paris)[6]. Une étiquette relative à la classe morpho-syntaxique a de plus été attachée aux lemmes. Les textes sont traités sous le logiciel UNITEX[7]. Des grammaires locales de désambiguïsation y ont été adaptées ; la levée des ambiguïtés lexicales, étape laborieuse de la lemmatisation, sera désormais partiellement automatisée (Kindt, 2003b : 10-12).

En dehors du cadre de la lemmatisation, les modes d'interrogation du dictionnaire ont été multipliés grâce à un encodage systématique des morphèmes constitutifs des lemmes. La présente contribution détermine la nature précise d'une telle opération, en présente les objectifs, et décrit les applications informatiques créées à cet effet.

1.2. *L'étiquetage des morphèmes constitutifs des lemmes du* D.A.G. *: nature et objectifs*

La lemmatisation consiste en un étiquetage lexical basé sur l'unité-mot. Les champs « lemmes » et « formes » de la base de données du *D.A.G.* peuvent faire l'objet d'interrogations. La formulation de ces requêtes se réduit toutefois à une chaîne de caractères, ce qui peut générer des réponses inadéquates. Une recherche sur <χειρ> « main », fournira ἀκροχειρίζω « toucher du bout des doigts », ἐπιχείρησις « entreprise », ou χειροποιητός « fait de main d'homme », réponses adéquates, mais aussi χείρων « pire », sans rapport avec χείρ « main ». Pour rendre l'exploration du *D.A.G.* plus efficace, il faut réduire le « bruit » amené par le manque de précision de ces requêtes. L'étiquetage lexical organisé au départ de l'unité-mot a donc été complété par un étiquetage des unités inférieures au mot, les morphèmes, et même, pour prendre en compte les thèmes, les combinaisons de morphèmes.

Une telle démarche requiert deux opérations concomitantes : le dénombrement des morphèmes et la mise en relation des lemmes présentant des morphèmes communs. Une base de données originale a donc été developpée. Au départ de celle-ci, une interrogation basée sur χείρ « main », fournit les septante-trois lemmes apparentés à ce mot (de ἀκρόχειρ, ἀκροχειρίζω, ἀνεπιχείρητος, ἀντίχειρ, ἀντιχειροτονέω, ἀπαρεγχείρητος, etc., à χειρώναξ). De plus, comme la nature des morphèmes et des combinaisons de morphèmes est annotée, l'interrogation peut être complétée par une requête touchant n'importe quel élément constitutif des mots, le suffixe -ιζε/ο-, par exemple, formant les verbes dénominatifs en -ίζ-ω/-ο-μαι-. La réponse affiche alors cent douze lemmes (depuis ἀκροχειρίζω, ἐγχειρίζω, ἀναλογίζομαι, διαλογίζομαι, etc., jusqu'à ὑπνίζω) (Gérard et Kindt, 2003). Ces matériaux et les interfaces

[4] Sur toutes ces notions, cf. Kindt (2003a : 1-19).

[5] Date de la première mention du *D.A.G.* dans Denis (1987 : XII).

[6] Le *D.A.G.* devient ainsi un DAG_DELAF ; sur les dictionnaires DELA, cf. Courtois (1990 : 11-22). Cf. aussi le site du Laboratoire d'Informatique Linguistique de l'Université de Marne-la-Vallée à l'adresse http://infolingu.univ-mlv.fr.

[7] Sur UNITEX, cf. http://www-igm.univ-mlv.fr.

de saisie et d'interrogation qui permettent de les manipuler sont rassemblés dans une nouvelle base de données baptisée *Dictionnaire Dérivationnel Grec* (*D.D.G.*). À ce stade, l'étiquetage est manuel : le travail progresse lemme après lemme, les grandes familles de mots appartenant à un même champ dérivationnel étant privilégiées. À l'avenir, l'analyse semi-automatique des constituants morphologiques des lemmes nouveaux est envisageable.

1.3. *Une idée ancienne, des moyens nouveaux*

Le *D.D.G.* est un dictionnaire affranchi de l'ordre alphabétique. Malgré son intérêt pratique, ce mode de classement traditionnel n'a pourtant aucun fondement linguistique. Il présente de plus un effet pervers qui est d'éloigner les uns des autres un grand nombre de mots linguistiquement apparentés. Regrouper les mots issus d'un même champ dérivationnel n'est pas une idée neuve. Dans son article intitulé *Towards an Electronic Greek Historical Lexicon* (1994), W.A. Johnson « rêve » d'une telle réalisation que les moyens informatiques contemporains rendaient déjà envisageable (Johnson, 1994). Prenant pour exemple le nom μανία et les adverbes en -κως, il illustre l'intérêt que présenterait un inventaire global des mots dérivés d'une même base et mis en relation avec la chronologie des sources dans lesquelles ces créations lexicales apparaissent pour la première fois. Il cite également le *Thesaurus Graecae Linguae* de H. Estienne, car la nomenclature de l'édition originale de ce monument de la lexicographie (1572) était structurée selon les racines dont les mots grecs dérivent, ἐπεισβαίνω figurant ainsi à la suite d'autres dérivés et composés de βαίνω, et non à la place que lui aurait assignée le respect de l'ordre alphabétique *stricto sensu* [8].

Trois facteurs ont facilité l'opération d'étiquetage des morphèmes. Le premier tient au fait que le *D.A.G.* était déjà constitué ; l'ensemble de ses lemmes propose un échantillon, incomplet, certes, au regard de la totalité du lexique grec, mais déjà représentatif. Le deuxième tient aux morphèmes qui relèvent, quant à eux, d'une série limitée d'éléments identifiables et distinctifs organisés en système, un ensemble d'éléments « nécessairement présents »[9] dans l'échantillon de référence. Le dernier facteur tient à l'essor de l'outil informatique ; une organisation appropiée de la base de données permet de fusionner en une seule étape la saisie des éléments morphologiques pertinents et leur mise en relation avec les lemmes dans lesquels ils apparaissent. Le travail présenté ici dépasse le rêve de W.A. Johnson car il suffira, à l'avenir, d'appliquer le *D.D.G.* à un *corpus* textuel pour dégager les procédés de formations lexicales qui s'y actualisent. Il dépasse aussi la réalisation de H. Estienne car il est désormais possible de classer les mots grecs non seulement selon leur racine, leur radical ou leur thème, mais aussi selon chacun des morphèmes (ou des combinaisons de morphèmes) qui les constituent.

2. Le *D.D.G.*

2.1. *Structure*

Les matériaux lexicaux présents dans le *D.D.G.* sont, à l'instar de ceux du *D.A.G.*, structurés en une base de données relationnelle. Celle-ci renferme plusieurs tables. La majorité de celles-ci rassemble les données morphologiques permettant de décrire les lemmes, c'est-à-dire les morphèmes *proprie dictu* (radicaux, racines indo-européennes, suffixes, préfixes, préverbes), et les combinaisons de morphèmes (thèmes). Une table rappelle les lemmes du *D.A.G.*

[8] H. Estienne s'explique lui-même sur la conception du *Thesaurus Graecae Linguae*, cf. l'édition récente des préfaces du lexicographe dans Kecskeméti *et al.* (2003), et spécialement, pour le regroupement des mots par racine, pp. 250-251 et 293-294.

[9] Expression reprise à D. et P. Corbin dans Corbin et Corbin (1991 : 147).

Une dernière table établit les relations entre les lemmes et les morphèmes. Une interface appelée *Interface de Caractérisation Morphologique des Lemmes* (*I.C.M.L.*) permet d'encoder ou d'afficher les données du *D.D.G.* Deux interfaces d'interrogation permettent d'en explorer le contenu.

2.2. *Formulaire d'étiquetage des morphèmes*

Le formulaire *I.C.M.L.* est utilisé pour identifier et encoder les morphèmes et les combinaisons de morphèmes des lemmes choisis par l'utilisateur, qu'il sagisse d'un mot simple, d'un dérivé ou d'un composé.

Pour le dérivé ἐπανάβασις, par exemple (cf. fig. 1), les champs suivants sont affichés à l'écran : – le lemme en cours (ἐπανάβασις) ; – sa base dérivationnelle (ἐπαναβαίνε/ο-) ; – son thème (ἐπαναβασι-) ; – le suffixe attaché au thème (-τ/σι- (δόσις)) ; – la désinence du lemme (-ς) ; – la nature du lemme (nom commun). Les champs « radical » et « racine » restent vides car ils ont déjà été saisis dans le formulaire correspondant au lemme simple βαίνω, respectivement βα(ν) et *g^{w}em-/g^{w}eh_2. Ces deux informations, communes à plusieurs lemmes apparentés, ne sont donc encodées que sous le formulaire du mot simple. La structure même de la base de données permet ensuite, lors d'une interrogation, de répercuter ses informations sur tous les lemmes concernés.

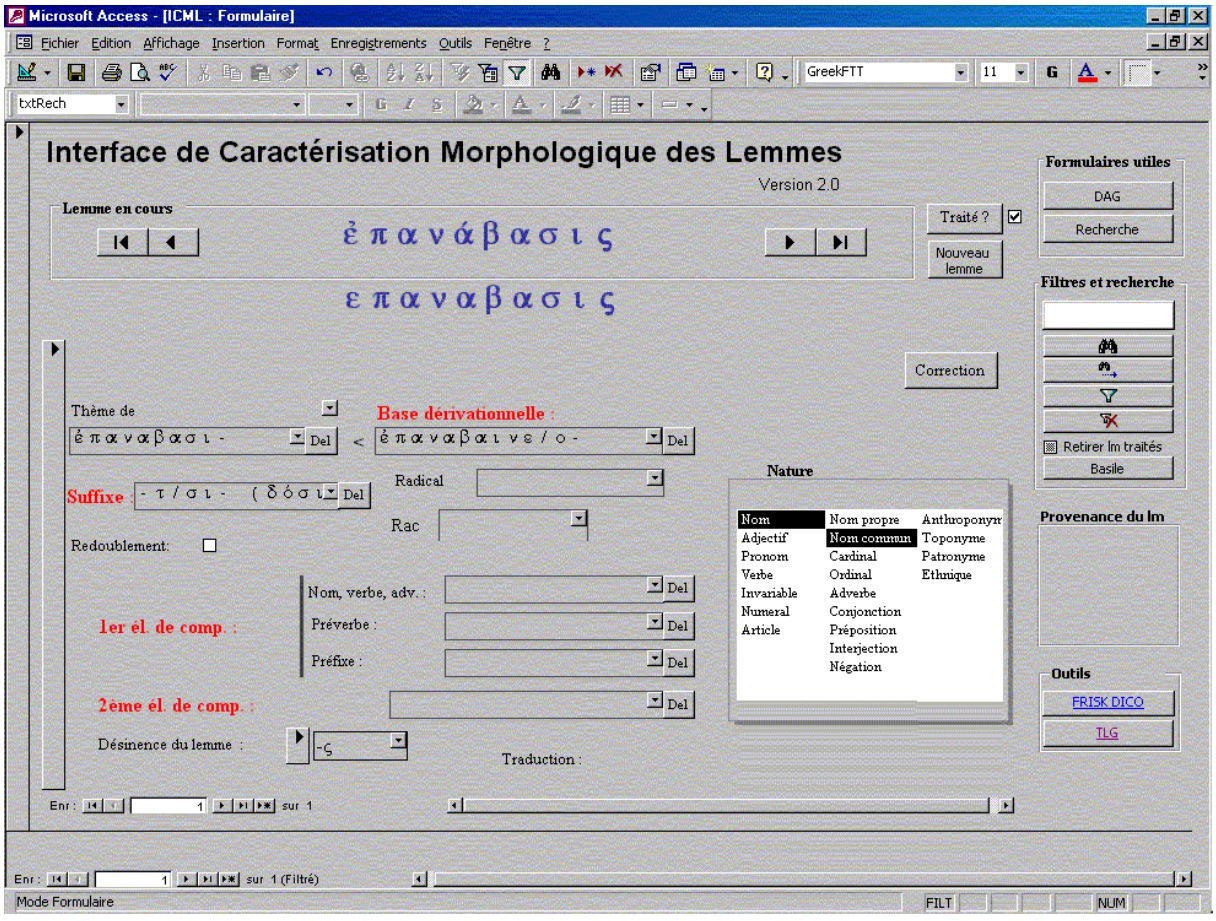


Figure 1. Formulaire d'étiquetage des morphèmes pour le lemme ἐπανάβασις.

Pour le composé φιλάγαθος, le formulaire indique, outre le lemme : – son thème (φιλα-γαθε/ο-) ; – le premier élément de composé (φιλε/ο-) ; – le second élément de composé (ἀγαθε/ο-) ; – la désinence du lemme (-ς) ; – la nature du lemme : adjectif.

Certaines entrées du *D.D.G.* sont précédées de l'arobase (par exemple, @ἀκρόβατος) indiquant que le mot est bien attesté dans les sources — ce qui établit son existence dans la langue —, mais qu'il ne figure pas dans la nomenclature du *D.A.G.* et est postulé afin d'expliquer les lemmes qui en dérivent (tel ἀκροβατέω). L'astérisque marque les entrées du *D.D.G.* qui, quoique absentes du *D.A.G.* et des sources disponibles, doivent être postulées au niveau de la langue pour justifier d'autres formations lexicales qui en dérivent et qui, elles, sont présentes dans le *D.A.G.* (par ex. *σκληροφάγος nécessaire pour décrire le lemme σκληροφαγία).

D'autres boutons du formulaire permettent d'appliquer des filtres, d'activer des raccourcis vers le *D.A.G.* lui-même, vers les interfaces d'interrogation et vers différents outils lexicographiques.

2.3. *Formulaire d'étiquetage de la classe morpho-syntaxique des lemmes*

Les lemmes ont reçu une étiquette identifiant leur classe morpho-syntaxique (cf. fig. 3). Les classes retenues sont les suivantes : nom commun ; nom propre (anthroponymique, patronymique, toponymique, ethnique) ; adjectif ; verbe ; pronom ; article ; muméraux (cardinaux, ordinaux) ; invariable (adverbe, conjonction, interjection, négation, préposition).

2.4. *Interfaces d'interrogation*

2.4.1. *Interface de recherche à critères multiples*

L'« interface de recherche à critères multiples » sélectionne des éléments et fournit les lemmes dans lesquels ceux-ci s'actualisent. L'utilisateur détermine d'abord le type d'élément (thème, suffixe, préverbe, préfixe ou radical). Il sélectionne ensuite un des éléments proposés dans une listbox et lance la requête. Le résultat s'affiche sous forme de liste. Quand une recherche est basée sur un thème, une zone permet de préciser s'il s'agit d'une base dérivationnelle, d'un premier ou d'un second élément de composé. Une autre zone permet de sélectionner des « critères supplémentaires » : la classe morpho-syntaxique des lemmes attendus par l'utilisateur et la nature de la base dérivationnelle, du premier ou du second élément de composé. Il est ainsi possible de rechercher quels sont les lemmes adjectivaux (« critères supplémentaires ») construits sur le thème (« type d'élément ») βαινε/ο- (« élément » tiré de la liste des thèmes). La réponse fournit les lemmes ἄβατος, βατός, δύσβατος et @ἀκρόβατος. La recherche peut porter simultanément sur trois « types d'élement » reliés par des opérateurs booléens. Il est ainsi possible d'extraire du *D.D.G.* les lemmes construits sur le thème βαινε/ο- et comprenant le suffixe -τ/σι- (δόσις) : le mot βάσις s'affiche. En excluant le suffixe -τ/σι- (δόσις), dix-sept réponses s'affichent, de βῆμα et ἄβατος à βάθρα. En limitant la requête aux lemmes dans lesquels le thème βαινε/ο- apparaît comme second élément de composé, dix résultats apparaissent : ἄβατος, δύσβατος, etc. En sélectionnant le thème βαινε/ο- et le suffixe -μα(τ)-, le résultat est βῆμα.

2.4.2. *Interface de recherche par descendance*

L'« Interface de recherche par descendance » fait état du caractère récursif des mots grecs. La récursivité est la propriété par laquelle le lexique renouvelle son stock d'unités-mots par dérivations ou compositions successives. Une base dérivationnelle accrue d'un suffixe est ainsi à

l'origine d'une création lexicale qui, à son tour, sera susceptible de constituer une nouvelle base : βαίνω → καταβαίνω → συγκαταβαίνω → συγκατάβασις.

L'« Interface de recherche par descendance » permet à l'utilisateur de sélectionner une base dérivationnelle et lui fournit tous les lemmes qui y sont apparentés. Le résultat s'affiche sous la forme d'une « arborescence dérivationnelle » faisant état de la récursivité du thème retenu (cf. fig. 2).

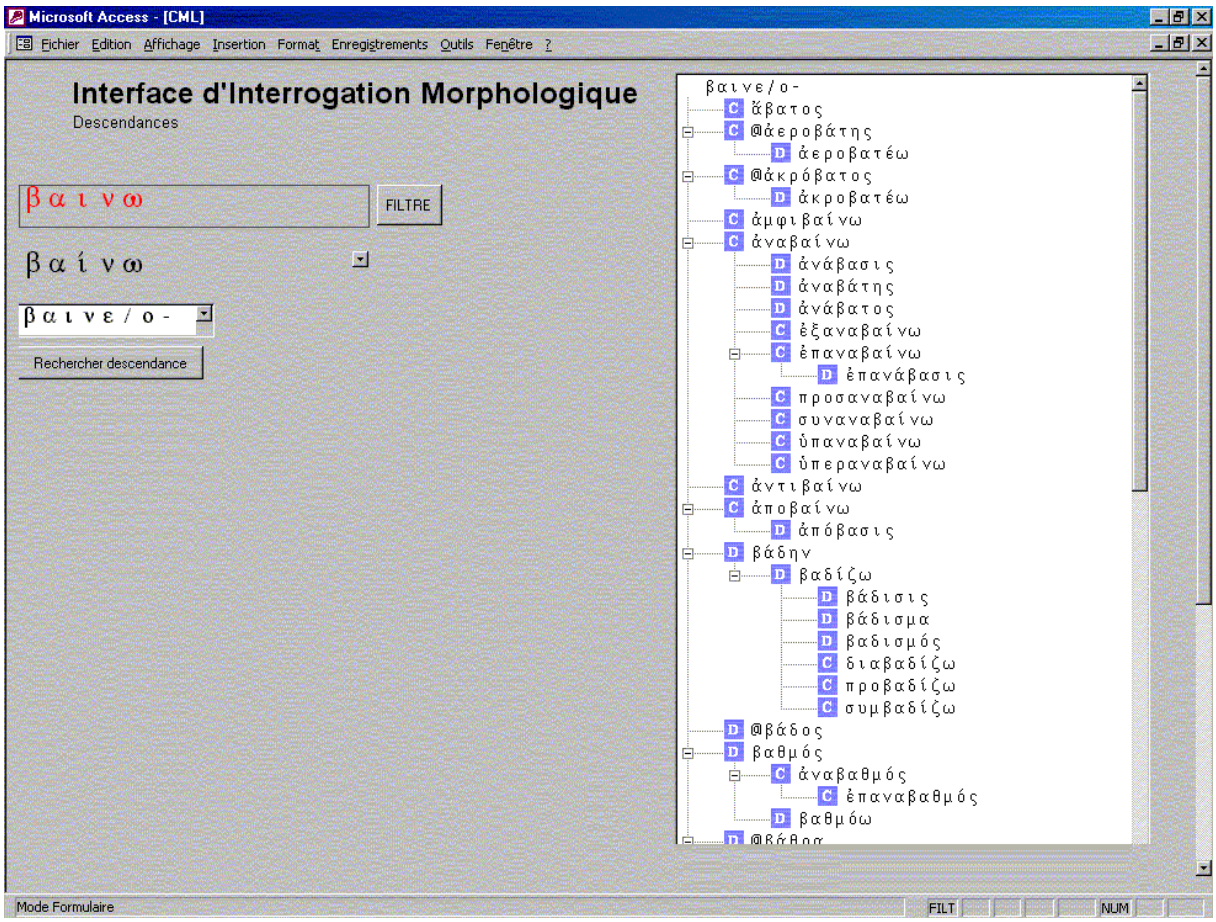


Figure 2. Arborescence dérivationnelle de βαίνω (extrait).

4. Conclusions

4.1. *État d'avancement du projet*

Le travail d'étiquetage des morphèmes s'étend à ce jour à 10.424 lemmes (soit 30,77% des matériaux lexicaux du *D.A.G.*). La classe morpho-syntaxique de 31.844 lemmes a été définie (cf. fig. 3). Ces tâches doivent se poursuivre. Dans un premier temps, les noms propres, les pronoms, l'article, les numéraux et les invariables ont été arbitrairement écartés de la caractérisation morphologique (ils représentent 5.430 lemmes, soit 16% du *D.A.G.*). Les anthroponymes grecs seront prioritairement pris en compte dans les analyses futures.

Ces données doivent être éprouvées et contôlées. Les interfaces d'interrogation permettront par exemple de vérifier la pertinence du dégroupement de certains suffixes homographes, tels -ιζε/ο- suffixe déverbatif (par exemple dans βαπτίζω, dont la base dérivationnelle est le verbe βάπτω), et -ιζε/ο- suffixe dénominatif (par exemple dans οἰκίζω, dont la base dérivationnelle

est le nom οἶκος). Les mêmes outils seront utilisés pour vérifier l'homogénéité des traitements. L'analyse des formations dites parasynthétiques sera par exemple réexaminée. Dans le *D.D.G.*, le lemme ἀδίδακτος dérive directement de διδάσκω et non de διδακτός. Est-ce que la même interprétation a été adoptée pour les autres formations analogues ?

D.A.G. 4.284.380 mots-occurrences			
formes	174.758	lemmes	33.874
D.D.G. (étiquetage des morphèmes) 10.424 lemmes traités			
racines	378	préfixes	10
radicaux	631	préverbes	12
suffixes	286	thèmes	10.779
D.D.G. (étiquetage des classes morpho-syntaxiques) 31.844 lemmes traités			
noms	12.740	articles	1
noms communs	8.972	numéraux	803
noms propres	3.762	cardinaux	703
anthroponymiques	1.842	ordinaux	100
patronymiques	17	invariables	823
ethniques	28	adverbes	645
toponymiques	1.704	conjonctions	47
adjectifs	8.730	négations	2
verbes	9.098	interjections	17
pronoms	43	prépositions	36

Figure 3. Données chiffrées des matériaux lexicaux du D.A.G. et du D.D.G.

4.2. *Perspectives ouvertes*

Tels qu'ils sont actuellement conçus, le *D.A.G.* et le *D.D.G.* permettent d'assurer la lemmatisation des sources et d'étudier les champs dérivationnels des mots enregistrés dans leur nomenclature commune. L'intérêt direct du *D.D.G.* est double. Lors de l'exercice de la lemmatisation, il offre les paradigmes utiles à l'étude de la productivité des morphèmes et de la régularité constructionnelle des mots rares rencontrés dans les sources. L'arborescence dérivationnelle permet d'établir facilement une liste, encore partielle, certes, mais fiable, des mots apparentés à un même terme, évitant aux utilisateurs un dépouillement fastidieux, et forcément imparfait, des dictionnaires traditionnels.

Les perspectives d'avenir pourraient orienter le projet vers la mise au point de programmes d'analyse. Un lemme inédit rencontré dans les sources ferait ainsi l'objet d'une caractérisation automatique de ses constituants. Les bases de données pourraient aussi accueillir des informations sémantiques permettant d'établir, par exemple, les relations de synonymie, d'antonymie, d'hyperonymie ou d'hyponymie qui traversent le lexique. Les applications présentes relient πύρ, ἐμπύριος et ζώπυρος, etc. Les outils futurs devraient rassembler ἐμπύριος « qui est en feu », αἴθοψ « qui est couleur de feu » et φλόγινος « enflammé, qui est couleur de feu ».

Enfin, les ressources et les outils conçus dans le cadre du projet seront progressivement mis à la disposition des utilisateurs sur la toile, *via* des interfaces web déjà opérationnelles en version expérimentale.

Références

Coulie B. (1996). La lemmatisation des textes grecs et byzantins : une approche particulière de la langue et des auteurs. *Byzantion*, vol. (66) : 35-54.

Coulie B. (2003). Thesaurus Patrum Graecorum. In Leemans J. (Ed.), *Corpus Christianorum 1953-2003. Xenium Natalicum. Fifty Years of Scholarly Editing* : 169-172.

Corbin D. et Corbin P. (1991). Vers le Dictionnaire dérivationnel du français. *Lexique*, vol. (10) : 147-161.

Courtois B. (1990). Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française*, vol. (87) : 11-22.

Denis A.-M. (1987). *Concordance des Pseudépigraphes d'Ancien Testament. Concordance. Corpus des textes. Indices*.

Gérard R. et Kindt B. (2003). Le Projet de recherche en lexicologie grecque à l'Institut orientaliste de l'Université catholique de Louvain. La collection du Thesaurus Patrum Graecorum et le Dictionnaire Automatique Grec. *Byzantion*, vol. (73) [à paraître].

Johnson W.A. (1994). Towards an Electronic Greek Historical Lexicon. *Emerita*, vol. (62) : 253-261.

Kecskeméti J., Boudou B. et Cazes H. (2003). *La France des humanistes. Henri II Estienne, éditeur et écrivain* (*Europa Humanistica*).

Kindt B. (2003a). *La lemmatisation automatisée au service d'une description lexicale du grec ancien. Propositions pour la formulation des lemmes du* Dictionnaire Automatique Grec *(D.A.G.). Rapport de recherche présenté en vue de l'obtention du D.E.A. en Philosophie et Lettres (ISLE 3DA)*, orientation « Philologie et histoire orientales ». Louvain-la-Neuve.

Kindt B. (2003b). Avancées dans le traitement automatique du grec ancien à l'U.C.L. L'analyse des textes au service d'une description lexicale de la langue. Une description lexicale de la langue au service de l'analyse des textes. In Labbé D. (Éd.), *Lexicometrica*, numéro spécial « Autour de la lemmatisation » : 1-17 (http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema1.htm).

(How) can causative constructions be predicted?

Gaëtanelle Gilquin[1], Éric Lecoutre[2]

[1] F.N.R.S. / CECL, LIGE, UCL – Place Blaise Pascal, 1 – 1348 Louvain-la-Neuve – Belgique
gilquin@lige.ucl.ac.be

[2] Institut de statistique, UCL – Voie du Roman Pays, 20 – 1348 Louvain-la-Neuve – Belgique
lecoutre@stat.ucl.ac.be

Abstract

This study investigates whether English causative constructions can be predicted, and if so, how. The technique used here is that of the decision tree, which emerges as the most powerful tool for our purposes. Some results are given and it is shown how these results can be applied, not only to a scientific description of causative constructions, but also, provided some changes are implemented, to the field of second or foreign language acquisition.

Résumé

Le but de cette étude est de déterminer si les constructions causatives anglaises peuvent être prédites, et si oui, comment. La technique utilisée à cet effet est l'arbre de décision, qui apparaît comme l'outil le plus approprié pour ce faire. On donnera quelques résultats, et on montrera que ces résultats peuvent être appliqués, non seulement à une description scientifique des constructions causatives, mais aussi, à condition que certains changements soient effectués, au domaine de l'acquisition d'une langue seconde ou étrangère.

Keywords: decision tree, prediction, causative constructions.

1. Introduction

As pointed out by Kemmer and Verhagen (1994:115), "[t]he grammar of causative constructions has inspired what is probably one of the most extensive literatures in modern Linguistics." This they explain by the "fascinating complexity of causatives both within particular languages and cross-linguistically," as well as the "tacit recognition by many linguists that an understanding of causatives is fundamental to an understanding of clause structure as a whole." Despite this complexity and centrality, and although causative constructions have been approached from a wide variety of perspectives – including (but not limited to) Generative Semantics and its classic derivation of *kill* from CAUSE BECOME NOT ALIVE (cf. McCawley, 1968), the universal-typological perspective (e.g. Comrie, 1976 or Wierzbicka, 1998) or the approach of functional grammar (see Dik, 1980) – it must be recognised with Altenberg and Granger (2001: 184) that "[i]t is very difficult to find a good description of the usage differences between [causative] verbs." Patchiness, focus on formal, rather than semantic aspects, and general lack of reliability are among the weaknesses of the descriptions of causative constructions found in the literature. This can be linked to the lack of empirical foundations of such descriptions for, as rightly emphasised by Kemmer and Verhagen (1994: 148), "a complete understanding of causatives must (…) take into account *empirically attested* semantic patternings (…), an area which has been relatively neglected in the literature so far" (emphasis added). The aim of the study presented here is precisely to take such empirically attested patternings into account, and see whether the choice of a particular causative verb can be predicted on this basis.

2. Material

Four English causative verbs have been investigated, viz. *cause*, *get*, *have* and *make*, as they are used in so-called periphrastic causative constructions (e.g. *He made her leave the room* or *I had the car repaired*). The data come from the British National Corpus (BNC) and consist of some 5 million words of spoken English (spontaneous conversations and broadcast discussions) and 5 million words of written English (academic prose from various fields). The number of causative constructions retrieved is shown in Table 1.

	n	%	/ 100,000 words
CAUSE	200	5.6%	2.04
GET	1,310	36.7%	13.36
HAVE	813	22.7%	8.29
MAKE	1,251	35.0%	12.76
TOTAL	3,574	100.0%	36.46

Table 1.Raw frequency (n), percentage and relative frequency per 100,000 words of the four causative verbs in the corpus

For each causative construction, a number of (syntactic and semantic) variables were examined and stored in a database, such as the tense of the verb, the animate or inanimate nature of the CAUSER,[1] or the form of the EFFECT (bare infinitive, *to*-infinitive, present participle or past participle). In total, some 40 variables were thus analysed.

3. Methodology

Several methods can be applied in order to predict the use of a particular word or construction, but they are not all equally appropriate to deal with causative verbs. Without any other available information than the observed frequencies, one can envisage two different ways to predict the causative verb, viz. raw prediction and naive prediction. The raw prediction always returns the same output, namely the most frequent verb, *get*, which leads to a prediction rate of 36.7% (i.e. the observed proportion of causative constructions that use *get*). The naive prediction consists in returning a random verb with a probability corresponding to its empirical proportion. The naive predictor results in a global rate of prediction of 30%.[2] Not only are these two methods simplistic, however, but they do not provide any explanation for why one form should be preferred rather than another. To provide such an explanation, one should investigate variables that might account for the choice of a particular causative. As can be expected from Kemmer and Verhagen's remark on the complexity of causatives, pinpointing a single variable will not do. Thus, Guierre's (1959: 126-7) claim that the choice between the different causatives can basically be reduced to a distinction between active sense ("acting," e.g. *She **made** her brother read the book*) and passive sense ("undergoing," e.g. *He **had** his watch fixed*) is obviously inaccurate, as most verbs can be used with both an active and passive meaning (compare *She made her brother read the book* with *She made her views*

[1] The CAUSER refers to the initiator of the causative process (e.g. ***He** had his house built in 1980*). The other terms used to designate the different elements of the causative construction are the CAUSEE, the entity which is changed or influenced by the CAUSER and carries out the EFFECT (e.g. *The teacher made **him** read the book*). The EFFECT represents the action thus performed (e.g. *He got the video **working***) and can be followed by an object, called the PATIENT (e.g. *The drought has caused millions of people to leave **their homes***).

[2] The formula to calculate this is: $\Sigma (pi)^2$, where pi = empirical proportion in the corpus of the i^{th} form.

known, or *He had his watch fixed* with *He had his class write an essay*). So clearly, several variables have to be taken into account. One way of doing this is by means of discriminant function analysis. This technique is used to determine which variables discriminate between a number of target categories (here, the categories of *cause*-, *get*-, *have*- and *make*-constructions). The variables thus emerging are the best predictors for the category. Although the technique can be used with excellent predictive power (cf. Gries's (2003) analysis of the dative alternation), it is important to bear in mind that it judges the variables individually, not in combination with one another. That this difference can be crucial appears from the following example. In our data, the CAUSER and CAUSEE with *make* are predominantly animate (although the former with a small majority), while the EFFECT tends to be non-volitional. Yet, the most frequent combination is not animate CAUSER + animate CAUSEE + non-volitional EFFECT, but **in**animate CAUSER + animate CAUSEE + non-volitional EFFECT (with a proportion of 28.40%, against 17.84% for the former combination). In other words, it turns out that the combination of the most frequent individual variables does not perforce correspond to the most frequent combination of variables. There are at least two techniques that examine the different variables in relation to one another, viz. neural networks and decision trees (see Berry and Linoff, 1997). Since decision trees, unlike neural networks, are directly expressed as understandable rules, they should be preferred in an analysis like ours which aims not only to predict the choice of one verb or another, but also (and above all) to explain this choice. The technique of decision trees has been applied successfully by Duhoux and Lecoutre (2003), who sought to establish the variables that influence the choice of verbal aspect in ancient Greek and managed to predict over 60% of the forms.

Decision trees can be built in the SAS (Statistical Analysis Software) System – "Enterprise Miner". At the root of the tree is the whole set of the training data. By means of the χ^2 test, all the variables are examined in order to determine the best candidate for a split, that is, the candidate that will divide the training data into two subgroups with the best predictive results. The same process is applied to the nodes thus obtained and is repeated until we reach a terminal node, called a "leaf". Each leaf corresponds to a particular prediction and a precise rule, which can be retrieved by looking at the path that connects the root to the leaf. The rules take the following form: "If X = a and Y = b, then Verb = v," followed by the percentage that v represents in the leaf, as well as the number of records that are concerned by the rule. The predictive power of the tree can be assessed by calculating the number of records that are correctly classified according to the tree. The reliability of the tree can then be checked by applying it to previously unseen data. If the tree performs as well with the new data as with the training data, the tree can be said to be reliable. Here, the sample has been divided into 60% of training data and 40% of validation data.

4. Results

Despite the caveat mentioned with respect to discriminant function analysis, this technique was tested against our data. While the verb *get* reaches a very good 88% of correctly predicted cases, the proportion falls to 67.5% with *have*, 36.5% with *cause* and under 25% with *make*. As a comparison, the decision tree with the same variables has a total predictive power of some 83%, taking eight leaves into account.[3]

However, an analysis of the results of the decision tree shows that it might be preferable to make a selection among the variables, rather than using them all, and in fact, to change the

[3] Beyond eight leaves, the results of the training data start to diverge from the validation data.

nature of the targets altogether. With no prior selection, the variable that best discriminates among the target classes is the form of the EFFECT, with the first rule stating that "If Form = Infinitive, then Verb = MAKE (93%)." This rule, which predicts 487 occurrences of *make* (that is, 90% of the total occurrences of *make* in the training data), tells us little more than an analysis of the frequency of the different complements with *make*. Moreover, this variable can be said to partially overlap with the target, since causative *make* is only possible with certain types of complements, including bare infinitive. Generally speaking, we do not primarily choose the verb *make* because we want to put the EFFECT in the bare infinitive, but we choose the bare infinitive because it is among the types of complements that are acceptable after *make*. Put differently, we tend to choose the causative verb first (according to rules that will have to be discovered), and then select one of the possible complements.[4] It was therefore decided to include the complement in the target and so turn the four targets into ten targets, that is each verb with the different complements it accepts, viz. CAUSETO (*cause* + *to*-infinitive), GETTO (*get* + *to*-infinitive), GETPP (*get* + past participle), GETPRP (*get* + present participle), HAVEINF (*have* + infinitive), HAVEPP (*have* + past participle), HAVEPRP (*have* + present participle), MAKEINF (*make* + infinitive), MAKETO (passive *make* + *to*-infinitive) and MAKEPP (*make* + past participle) (see Table 2 for the frequency of the different patterns).[5] This decision is supported by the analysis of the individual variables, which underlines the specificity of the different causative structures (e.g. large proportion of inanimate CAUSEES in present participle constructions with *get* and *have*, but not in infinitive constructions).

	n	%	/ 100,000 words
CAUSETO	200	5.6%	2.04
GETTO	366	10.2%	3.73
GETPRP	129	3.6%	1.32
GETPP	815	22.8%	8.31
HAVEINF	72	2.0%	0.73
HAVEPRP	70	2.0%	0.71
HAVEPP	671	18.8%	6.84
MAKEINF	1,120	31.3%	11.42
MAKETO	100	2.8%	1.02
MAKEPP	31	0.9%	0.32
TOTAL	3,574	100.0%	36.46

Table 2. Raw frequency (n), percentage and relative frequency per 100,000 words of the ten causative patterns in the corpus

[4] This is a simplification, to some extent, as some types of complements are not possible with all four causatives (cf. present participle, possible with *get* and *have*, but not with *cause* and *make*) and so might contribute to the choice of, say, one pair of verbs rather than another. However, the definition of the new targets will not ignore the parameter of the form of the EFFECT.

[5] Other types of complements are sometimes found in the corpus data, cf. (i) or (ii), but these are very marginal and so will not be taken into account here.

(i) What's **made** you *to think* of that? <BNC:S:KBB 612>

(ii) How do we **get** them *drink* more? <BNC:S:KBD 4671>

In addition, a number of variables were removed,[6] either because they had become redundant with the introduction of the new targets (e.g. voice of the EFFECT), or because they had not emerged in the decision tree built on the basis of all the variables (this is the case of the PATIENT, which, beyond its mere presence or absence, does not seem to influence the choice of the causative).[7] The remaining variables represent a total of 15.

The new tree has a total predictive power of some 73%. Granted, this is less good than the first decision tree, but still much better than a naive prediction (the naive predictor, it will be reminded, has a power of 30%). And in this new tree, there is no overlap between the targets and the variables. It is important to bear in mind, however, that not all targets are equally well predicted by the tree. In fact, there are two structures that are not predicted at all (at least at an "acceptable" level, i.e. at a level where the training data and validation data still coincide), namely *have* + present participle and *have* + infinitive. For the other targets, the percentage of correctly predicted cases ranges from 4.29% with *get* + present participle to 87% with *make* + infinitive and passive *make* + *to*-infinitive. The tree shows broad tendencies, concerning 300 or more forms, but also small "niches", which are applicable to just a few cases. Some of these should be discarded as irrelevant, but others can give us useful information about a particular structure. Thus, the rule "If PATIENT = no and CAUSEE = no, then Verb = MAKEINF (100%)," concerning only nine forms, points to particular constructions with *make* where neither the CAUSEE nor the PATIENT is expressed, cf.

(1) Well I have in the past given them fifty each for Christmas but this year they're going to have to **make** do with twenty. <BNC:S:KBF 4752>

(2) This is direct experience, but it is not drama -- not until there is some pretence involved, some symbolic representation, some intention to **make** believe. <BNC:W:AM6 126>

As for the general tendencies, they tell us, for example, that if the CAUSER is inanimate and the CAUSEE is mentioned, *make* + infinitive is used in 78% of the cases, which represents 521 forms, cf.

(3) And we also know that advertising reinforces smoking, it **makes** people think that smoking is okay. <BNC:S:FLM 346>

or that *have* and *get* + past participle are used in similar environments (viz. an animate CAUSER, a PATIENT, no CAUSEE and a volitional EFFECT), with the difference that *have* is chosen when the CAUSER is distinct from the (implied) CAUSEE (303 forms), while *get* is preferred when the CAUSER is co-referential with the CAUSEE or when the relation between the two participants is ambiguous (617 forms). Compare:

(4) She is always going to the hairdresser and **having** her hair frizzed <BNC:S:KCN 4895>

(5) But she couldn't **get** the car started this morning. <BNC:S:KBY 50>

We also learn that some structures are used in more specific environments than other structures, which are more "versatile." *Make* + infinitive is the structure that most often recurs as a leaf (6 leaves out of 14), which means that it can be used in many different contexts, includ-

[6] This, incidentally, facilitates the decision tree building, as it would be an extremely slow and (for the computer) tedious process to predict the use of ten targets on the basis of some 40 variables.

[7] In fact, the analysis of the individual variables reveals that the PATIENT, when expressed, presents almost identical characteristics with all four verbs (most notably, a predominantly inanimate nature, while the CAUSER and CAUSEE greatly differ in this respect).

ing the two contexts already alluded too, but also e.g. an animate CAUSER, a CAUSEE and a non-volitional EFFECT, as in:

(6) Now you've **made** me forget what I was gonna say <BNC:S:KP2 256>

By contrast, the use of *make* + past participle is very restricted, as all the forms correctly predicted (i.e. 11 forms out of 16) are of the same type, viz. an animate CAUSER, a non-volitional EFFECT, a PATIENT and no CAUSEE, e.g.

(7) The group who face the most difficulty getting adequately trained are women doctors, and they **made** their voices heard at the conference despite the organisers admitting with startling candour that they had not anticipated that this issue would arise. <BNC:W:CNA 215>

Now that we have a better idea of the kinds of results that are obtained by means of the decision tree, let us see how these results can be applied.

5. Applicability

5.1. *Predictive power*

Although the technique of decision trees is primarily meant to predict the membership of new cases, it should be noted that the immediate use of the predictive power of a tree is very restricted when it comes to linguistic phenomena. While it makes sense, say, to predict, for a new population, whether a person is a potential buyer of a particular product or not, or whether a credit card holder is likely to become an inactive customer (example given by Berry and Linoff, 1997: 268-273), it is of little use in itself to be able to predict the form that will be used in a given context. Except within the (necessarily artificial) frame of an exercise for learners, one would probably never be given a text from which all the causative verbs are missing and where the forms would have to be retrieved by means of an algorithm, for the simple reason that the verb is inseparable from the rest of the sentence. The (direct) usefulness of the predictive power of a decision tree in linguistics is therefore limited to the indication it gives of the reliability of the tree. However, decision trees become extremely powerful tools when they are applied in a more indirect way.

5.2. *Scientific description of the use of causative constructions*

What is useful in decision trees for our purposes is the set of rules that are automatically generated. Not only do they highlight the most relevant variables, but they also show which combinations of variables lead to the use of a particular verb. In other words, decision trees make it possible to describe the behaviour of causative constructions in real English and pinpoint the factors that influence the choice of one structure or another. As pointed out above, this is the sort of information that is most sorely lacking in the literature on causative constructions, although it is indispensable in order to make an informed choice, one that will sound natural and authentic.

This description, however, should be supplemented by other aspects, which are less adequately dealt with by means of a decision tree, but are nonetheless relevant to a description of the phenomenon under investigation. This is the case with the stylistic and lexical aspects of causative constructions. Introducing the speech/writing dichotomy in the decision tree is feasible, although it does not give rise to any significant improvement in the predictive power (74% of correctly predicted cases instead of 73%). As for genres, their great diversity makes the tree rather confusing, with rules such as "If genre is one of: conv disc hum med pol soc, then…" The whole stylistic aspect is therefore best described as an independent level, on a

par with the syntactic and semantic aspects described by the decision tree. The same applies to lexis. A collocational analysis of causative constructions reveals that each structure keeps strong preferential lexical company. Thus, the past participle with *make* frequently takes the form of *known* (18 occurrences in our data, that is 60% of all past participle constructions with *make*), *felt* (5 occurrences) or *understood* (2 occurrences), while with *get* and *have*, by far the most common EFFECT is *do* (almost 25% of all the EFFECTS with each causative). In addition, *get* frequently co-occurs with words referring to some sort of difficulty or effort, as in:

(8) We're not picking on you we're just trying to **get** a conversation going here like, but it's very hard when you just sit there and say nothing. <BNC:S:KCX 6634>

(9) Attempts to **get** parents to reduce calorific intake, if the cause of the obesity is psychogenic, are doomed to failure; possibly this is why these families are often so difficult to treat. <BNC:W:CGT 1461>

Again, introducing such elements would make the decision tree much more complex (also for the computer), resulting in rules of the type "If EFFECT is one of: *know feel understand*, then…" or "If context is one of: effort difficulty, then… ." Therefore, it is perhaps preferable, for the sake of simplicity, to treat the lexical aspect of causative constructions separately from the decision tree.

5.3. *Pedagogical grammar?*

By alluding to simplicity in the preceding section, we are already moving towards another field of application for decision trees, namely pedagogy and more particularly second and foreign language acquisition. That English causative constructions are difficult to use for non-native speakers appears from various studies (e.g. Wong, 1983; Liu and Shaw, 2001 for Chinese learners; Altenberg and Granger 2001 for Swedish and French-speaking learners; or Helms-Park, 2001 for speakers of Hindi-Urdu and Vietnamese). And pedagogical grammars are of little help here, as they suffer from the same problems as the rest of the literature, namely incomprehensiveness and unreliability.

Although a decision tree built on the basis of authentic data can certainly provide learners with practical information to choose the most suitable causative structure in a given environment, it seems to us that the tree has to be adapted before it can be of real use to them. First, only the most relevant and basic variables should be taken into account, those which are at the learner's immediate disposal **before** constructing the causative structure (number of participants involved, animate or inanimate nature of the entities, degree of volitionality of the EFFECT). Second, it would be necessary to get rid of some of the uninteresting rules that account for only a few records (see above), as well as rules which make little sense (at least to students) such as "If CerEv is one of: no yes, then…," where the only possibility that is discarded is a "non-applicable" value for the nature of the CAUSER (event or not). Then, the causative structures that are not predicted by the automatic tree should somehow be taken into account in a "pedagogical decision tree." This would involve allowing for more than one structure per leaf, with a note explaining the difference between the alternatives. The decision as to which additional structures to mention could be based on the proportion this structure represents, not vis-à-vis all the records classified by the rule in question, but vis-à-vis all the occurrences of this particular structure. To illustrate this, let us take the example of present participle constructions with *get*. The rule "If CAUSER = yes and CAUSEE = yes and PATIENT = no and CAUSER = animate and CAUSEE = inanimate, then…" provides a majority of MAKEINF records (116 out of 247, that is 46.96%). In an automatic decision tree, the leaf would therefore be labelled as MAKEINF. However, the records also include 64 instances of *get* + present

participle constructions. Although this is less than MAKEINF, these instances of GETPRP represent 49.61% of all the *get* + present participle constructions present in the corpus (64 constructions out of 129). So it is well worth mentioning in a pedagogical tree, since this rule shows one of the typical environments in which GETPRP occurs. More generally, one should systematise the tree as much as possible, so as to make it understandable to learners and easy to remember, even if it entails a decrease in precision. The resulting tree will thus have a weaker predictive power, but will be more effective in terms of second or foreign language acquisition.

6. Conclusion

Despite the complexity of causative constructions and the vagueness with which they tend to be described in the literature, it has been shown that the choice of a given causative verb and structure can be predicted to a certain extent by means of a decision tree. The main use of such a tree is that it emphasises the most relevant variables influencing this choice and, above all, the combinations of variables that lead to a particular choice. For such a tree to be usable in the field of second or foreign language acquisition, however, human intervention is needed in order to make the tree more economical and systematic.

References

Altenberg B. and Granger S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics*, vol. (22): 173-194.

Berry M.J.A. and Linoff G. (1997). *Data Mining Techniques for Marketing, Sales, and Customer Support.* Wiley and Sons.

Comrie B. (1976). The syntax of causative constructions: cross-language similarities and divergences. In Shibatani M. (Ed.), *The Grammar of Causative Constructions*. Academic Press: 261-312.

Dik S.C. (1980). The Dutch causative construction. In *Studies in Functional Grammar*. Academic Press: 53-89.

Duhoux Y. and Lecoutre E. (2003). La prédiction de l'aspect verbal en grec ancien. In *Proceedings from Convegno Internazionale di Linguistica Greca*.

Gries S.Th. (2003). Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics*, vol. (1): 1-28.

Guierre L. (1959). *The right Word … Le Mot juste … and the right Sound*. Librairie Vuibert.

Helms-Park R. (2001). Evidence of lexical transfer in learner syntax. The acquisition of English causatives by speakers of Hindi-Urdu and Vietnamese. *Studies in Second Language Acquisition*, vol. (23): 71-102.

Kemmer S. and Verhagen A. (1994). The grammar of causatives and the conceptual structure of events. *Cognitive Linguistics*, vol. (5): 115-156.

Liu E.T.K. and Shaw P.M. (2001). Investigating learner vocabulary: a possible approach to looking at EFL/ESL learners' qualitative knowledge of the word. *International Review of Applied Linguistics in Language Teaching (IRAL)*, vol. (39): 171-194.

McCawley J.D. (1968). Lexical insertion in a transformational grammar without deep structure. In Darden B.J., Bailey Ch.-J.N. and Davison A. (Eds), *Papers from the Fourth Regional Meeting of the Chicago Linguistic Society*, vol. (4): 71-80.

Wierzbicka A. (1998). The semantics of English causative constructions in a universal-typological perspective. In Tomasello M. (Ed.), *The New Psychology of Language. Cognitive and Functional Approaches to Language Structure*. Lawrence Erlbaum Associates Publishers: 113-153.

Wong S.C. (1983). Overproduction, underlexicalisation, and unidiomatic usage in the 'make' causatives of Chinese speakers. *Language Learning and Communication*, vol. (2): 151-163.

Il lessico della guerra nei newsgroups della categoria it.politica durante la guerra in Iraq

Luca Giuliano

Dipartimento di Contabilità nazionale e analisi dei processi sociali –
Università degli studi di Roma "La Sapienza"
luca.giuliano@uniroma1.it

Abstract

How have reacted the Italian newsgroups to the war in Iraq? During the 28 days of "declared" war, what have been the main political and ideological guidelines?
The monitoring of the 8 newsgroups of the it.politica category has permitted to explore this topic. The more frequent non empty word in the corpus is <guerra> (15.605 occurrences on 5.220.932 occurrences in total). This paper intends to analyze the use of this form with a minimal loss of information and a meaningful gain of thematic deepening in a very "noisy" and "dirty" source. For this purpose, the analysis is performed on a sub-text extracted by concordances with keywords <guerra>.
In complex, there is a climate of general opposition to the war, although with various ways and reasonings. It is not always possible to give a description of the newgroups based on the political choice of the users. They search the argument and the challenge with the political opponents.

Riassunto

Come hanno reagito i newsgroups italiani alla guerra in Iraq? Durante i 28 giorni di guerra "dichiarata" quali sono stati gli orientamenti politici ed ideologici prevalenti?
Il monitoraggio degli 8 newsgroups della categoria it.politica ha permesso di esplorare questo tema a partire dalla forma grafica <guerra>, la più frequente nel corpus (escluse le parole vuote) con 15.605 occorrenze su 5.220.932 occorrenze in totale. L'obiettivo è di minimizzare la perdita di informazione ma, nel contempo, ottenere un guadagno significativo di approfondimento tematico in una fonte di informazione nota per essere molto "rumorosa" e "sporca". A questo scopo, l'analisi viene condotta su un sub-testo estratto dal corpus e composto dalle concordanze della forma <guerra>.
In complesso, ne emerge un clima di opposizione generalizzata alla guerra, sebbene con modalità e argomentazioni diverse. Non sempre è possibile dare una caratterizzazione precisa del newsgroup in base all'orientamento politico che lo identifica nominalmente. Gli utenti dei newsgroups cercano la discussione e il confronto polemico con gli avversari politici.

Keywords : newsgroups, textual analysis, correspondence analysis, politics, war.

1. Introduzione

La guerra in Iraq è iniziata ufficialmente alle 4:14 (ora italiana) del 20 marzo 2003, con l'annuncio del presidente degli Stati Uniti, George Bush: "E' cominciata la guerra di liberazione dell'Iraq". In effetti già poco meno di un'ora prima, alle 3:35 (5:35 ora locale), i primi missili americani cadevano su Bagdad, mentre i marines e i soldati britannici entravano in territorio iracheno Il 2 aprile le truppe americane arrivavano alla periferia della capitale. Il 9 aprile, alle 16:49, le televisioni di tutto il mondo mostravano l'abbattimento della statua di Saddam Hussein in piazza al Ferdous. Da questo momento in poi l'Iraq passava progressivamente sotto il controllo anglo-americano. Il 16 aprile le operazioni di guerra potevano considerarsi finite con la richiesta, inoltrata dalla Casa Bianca all'ONU, di togliere l'embargo sull'Iraq.

I testi di cui trattiamo in questa comunicazione provengono da newsgroups di argomento politico. I newsgroups rappresentano una parte consistente della messaggeria elettronica asincrona. Sono più di 25.000 e generano circa 1.000.000 di messaggi al giorno. La loro rilevanza nei processi di comunicazione, non solo all'interno della comunicazione mediata dal computer, è stata ampiamente documentata (Smith, 1999; Choi e Danowski, 2003). Il sistema Usenet è facilmente accessibile ed è conosciuto in tutto il mondo dal 1979. In questi anni si è sviluppato fino a coprire un arco completo di argomenti e con diramazioni in moltissime aree linguistiche e culturali. Per la loro struttura sociale fondamentalmente "anarchica", i newsgroups sono particolarmente adatti per l'analisi delle correnti endogene nei conflitti sociali emergenti (Stubbs, 1998; Giuliano, 2002; Smith, 2002). L'archivio dei messaggi con interfaccia web, che è stato per diverso tempo gestito da Deja News, recentemente è stato acquistato da Google.

I newsgroups coinvolgono milioni di persone con una complessità di temi e di interazioni sociali che non si possono analizzare con gli strumenti attuali di navigazione e di "information retrieval" (Whittaker *et al.*, 1998).

L'analisi automatica dei dati testuali può offrire un contributo rilevante per l'individuazione dei nuclei di significato in una massa gigantesca di informazioni. In questa comunicazione si prendono in esame i newsgroups italiani durante la guerra in Iraq, la prima vera e propria guerra nell'era di Internet.

2. Corpus dei messaggi analizzati ed estrazione delle concordanze

La seconda guerra del Golfo ha avuto una durata di 28 giorni. Il periodo di riferimento per il prelevamento dei messaggi dai newsgroups della categoria it.politica va dal 18 marzo al 18 aprile.

Newsgroups	N. messaggi	Media msg al giorno	Identificato nel testo con:
it.politica	13.931	435,34	POLITICA
it politica.internazionale	6.194	193,56	INTERNAZIONALE
it.politica.pds	2.981	136,25	PDS
it.politica.lega-nord	3.809	119,03	LEGA-NORD
it.politica.polo	4.496	112,53	POLO
it.politica.rifondazione	2.728	90,91	RIFONDAZIONE
it.politica.destra	686	21,44	DESTRA
it.politica.cattolici	667	20,84	CATTOLICI
it.politica.ulivo	*349*	*11,40*	*scartato*
it.politica.libertaria	*245*	*7,65*	*scartato*

Tavola 1. Newsgroups della categoria it.politica secondo i messaggi nel periodo considerato: 18 marzo – 18 aprile 2003

I newsgroups della categoria it.politica sono 10, ma it.politica.libertaria e it.politica.ulivo sono stati scartati dalla rilevazione per il numero troppo esiguo di messaggi rispetto a una soglia arbitraria ma "ragionevole" di 20 messaggi in media al giorno.

Il corpus è costituito così da un file di 33Mb denominato IRAQ28 sottoposto a normalizzazione completa in TALTAC 1.6 con le seguenti caratteristiche lessicometriche:

Occorrenze	N	5.220.932
Forme grafiche	V	179.112
Type/Token ratio	(V/N)*100	3,43
Percentuale di *hapax*	(V1/V)*100	46,85
Frequenza media generale	N/V	29,49

Tavola 2. Misure lessicometriche del corpus IRAQ28

Sebbene già "ripulito" dell'*header* dei messaggi (l'intestazione che contiene le indicazioni di mittente, destinatario, data di invio, soggetto, server di posta, ecc.), il corpus si presenta con un'alta componente di "rumore" dovuta a varie fonti: marcatori HTML, indirizzi internet e di email, caratteri non riconosciuti dal server, errori ortografici, "firme" dei mittenti che utilizzano disegni in caratteri ASCII, emoticon, parole e interiezioni gergali, quoting, copia-e-incolla da pagine web, sovrapposizione di idiomi diversi, ecc.).

L'alta quota di rumore contenuta nel corpus è misurabile approssimativamente attraverso le forme grafiche non riconosciute dal tagging grammaticale di TALTAC. Le forme grafiche non riconosciute sono 76.984 (pari al 42,98% delle forme grafiche distinte contro una media del 5% nei testi letterari e giornalistici). Nonostante questo, per l'analisi del lessico e per l'analisi multidimensionale del testo, il rumore può essere trascurato. Infatti la quota maggiore di forme non riconosciute (26,67% delle forme grafiche) si trova tra gli hapax (le forme grafiche di occorrenza 1). Assumendo come riferimento per l'analisi la soglia di frequenza 17 consigliata da TALTAC, la percentuale di rumore relativa al testo da analizzare scende drasticamente al 5,04% delle occorrenze, con una copertura del testo del 90,98%.

La fascia di alta frequenza contiene 133 forme grafiche, tra le quali, se escludiamo le forme grammaticali e le forme banali che derivano dal linguaggio di Internet, troviamo le parole chiave principali che dimostrano una reattività eccezionale degli utenti di questi newsgroups agli eventi in corso.

Forme	Occorrenze	Rango
guerra	15.605	35
Iraq	9.011	56
Saddam	6.208	81
mondo	5.413	93
Bush	5.381	97
Governo	4.587	117
USA	4.276	124
Italia	4.237	126
Americani	4.048	130
Pace	4.023	131

Tavola 3. Forme grafiche principali appartenenti alla fascia di alta frequenza nel corpus IRAQ28

Tra le forme più significative,"guerra" è la più frequente. Seguono i principali riferimenti al contesto (Iraq, Saddam, Bush, USA, americani, pace). Pertanto, avendo come obiettivo di analizzare il tema principale del corpus con il minimo di trattamento del testo e con procedure il più possibile standardizzate e automatiche si è proceduto nel seguente modo:

- Estrazione delle concordanze della forma grafica "guerra/guerre" (Lexico 3).
- Estrazione dei segmenti con soglia 7 (SPAD_5.0).

- Costruzione della tabella lessicale a soglia 20 ed eliminazione dei segmenti ripetuti vuoti, banali e/o ridondanti.
- Analisi delle corrispondenze sulla tabella lessicale dei segmenti ripetuti per newsgroups (SPAD 5.0).
- Analisi delle corrispondenze sulla tabella lessicale a soglia 50 delle forme per concordanze (*mots-réponses*) con SPAD-T 1.5.
- Classificazione delle concordanze sugli assi fattoriali individuati con SPAD-T 1.5.

L'estrazione delle concordanze (70 caratteri prima e 70 caratteri dopo la forma pivot <guerra>) ha portato alla individuazione di 18.698 stringhe che costituiscono il sub-corpus GUERRA-IRAQ.

Occorrenze	N	474.476
Forme grafiche	V	32.945
Type/Token ratio	(V/N)*100	6,94
Percentuale di *hapax*	(V1/V)*100	45,18
Frequenza media generale	N/V	14,40

Tavola 4. Misure lessicometriche del sub-corpus delle concordanze GUERRA-IRAQ

Il rumore contenuto nel sub-corpus è notevolmente ridotto. Le forme grafiche non riconosciute dal tagging grammaticale di TALTAC sono 8.659 (pari al 26,28% delle forme grafiche distinte, ivi incluso il taglio di parole dovuto a +/- 70 caratteri intorno al pivot); alla soglia prescelta per l'analisi delle corrispondenze (soglia 20, copertura del testo 80%) le forme non riconosciute rappresentano solo il 3,69% delle forme grafiche.

Per analizzare il contenuto dei messaggi si è scelto di fare affidamento sulla individuazione dei temi in discussione attraverso la frequenza dei nuclei semantici rappresentati dai segmenti ripetuti nelle concordanze.

Nei newsgroups analizzati il clima di opposizione verso la guerra è prevalente. Il segmento ripetuto più frequente, a parte i segmenti di carattere grammaticale, è <contro la guerra> con 780 occorrenze.

I newsgroups si differenziano per il riferimento politico, salvo i due newsgroups generali ("politica" e "internazionale") che dovrebbero rispettivamente essere dedicati alla politica italiana e alla politica internazionale. Gli altri indicano chiaramente una scelta ideologica: it.politica.destra accoglie le discussioni intorno alla cultura di destra estrema; it.politica.leganord tratta argomenti relativi al federalismo; it.politica.polo è dedicato ai partiti di maggioranza della "Casa delle Libertà"; it.politica.pds è dedicato al partito dei Democratici di Sinistra; it.politica.rifondazione tratta argomenti che riguardano la sinistra di Rifondazione Comunista. Tuttavia gli utenti dei newsgroups non si differenziano nettamente secondo questi schieramenti. E' piuttosto diffusa la consuetudine di inviare una parte dei messaggi in tutti i newsgroups contemporaneamente (*crosspost*) sebbene sia vietato dalla *netiquette*. Gli utenti amano dibattere dei loro argomenti preferiti con gli "avversari". Accade piuttosto spesso che alcuni utenti si lascino andare a vere e proprie provocazioni, fingendosi appartenenti alla parte politica avversa o inondando il newsgroups di insulti.

3. Analisi dei segmenti ripetuti

L'analisi delle corrispondenze binarie segmenti/testi con soglia 20 effettuata con il programma SPAD 5.0 permette di individuare con qualche difficoltà i gruppi tematici caratteristici dei diversi newsgroups.

Numero	Autovalori	% di inerizia	% Cumulata di inerzia
1	0.0685	22.37	22.37
2	0.0587	19.17	41.54
3	0.0522	17.05	58.59
4	0.0415	13.57	72.16
5	0.0338	11.03	83.19
6	0.0287	9.36	92.55
7	0.0228	7.45	100.00

Tavola 5. Analisi delle corrispondenze dei segmenti ripetuti per testi: estrazione degli autovalori

Newsgroups	Coordinate		Contributi assoluti	
	F1	F2	F1	F2
CATTOLICI	0.31	0.86	6.9	59.8
DESTRA	-0.49	-0.02	13.1	0.0
INTERNAZIONALE	-0.32	0.15	39.2	10.7
POLITICA	0.25	-0.04	26.6	0.7
LEGA-NORD	0.12	-0.19	1.7	4.4
PDS	-0.11	-0.29	2.1	16.9
POLO	0.29	-0.04	9.7	0.2
RIFONDAZIONE	-0.08	-0.23	0.8	7.3

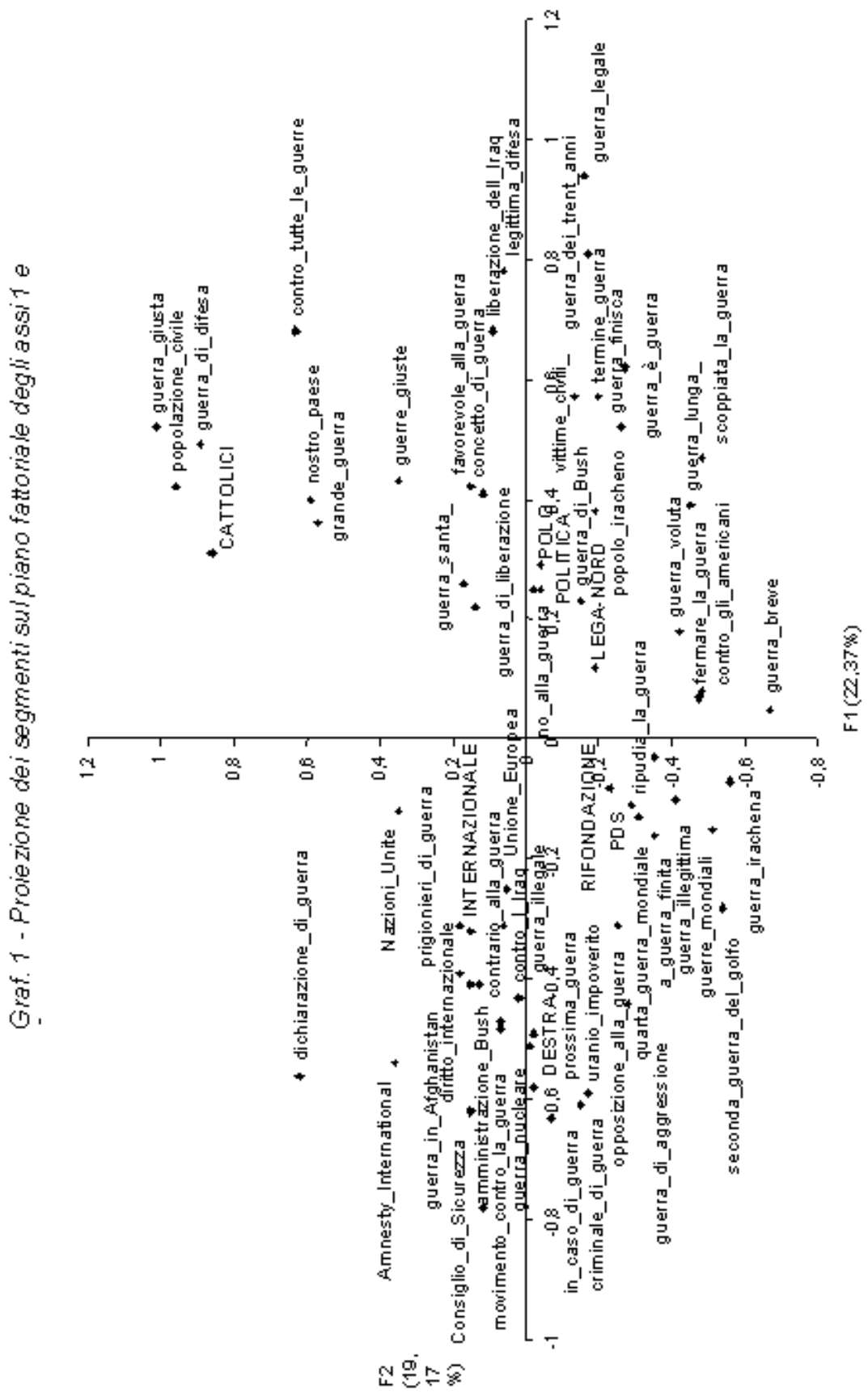
Tavola 6. Analisi delle corrispondenze dei segmenti ripetuti per newsgroups: coordinate e contributi assoluti delle frequenze attive

In questa comunicazione vengono presi in esame i primi due assi fattoriali che spiegano il 41,54% della variabilità complessiva. Le coordinate e i contributi assoluti delle frequenze attive permettono di individuare una diversificazione attesa sul primo fattore tra i due newsgroups principali, it.politica e it.politica.internazionale, intorno a un asse di politica interna e politica estera. Non sorprende la collocazione sul versante internazionale del newsgroup it.politica.destra che in passato ha ospitato con una certa intensità messaggi che denotano una forte adesione ai principi della destra estrema, antisionista ed estranea ai temi della politica nazionale.

Il secondo asse fattoriale si caratterizza per la contrapposizione tra it.politica.cattolici e it.politica.pds intorno a temi che fanno riferimento al dibattito sulla opposizione o no alla guerra e alle sue motivazioni. In tab. 7 sono riportati solo i segmenti ripetuti che presentano un contributo significativo per la formazione di almeno uno dei due fattori estratti.

SEGMENTI	Coordinate		Contributi assoluti	
	F1	F2	F1	F2
a guerra finita	-0,16	-0,35	0,1	0,8
amministrazione Bush	-0,41	0,13	1,3	0,2
Amnesty International	-0,54	0,36	1,3	0,7
concetto di guerra	0,41	0,12	0,7	0,1
Consiglio di Sicurezza	-0,62	0,15	2,6	0,2
contrario alla guerra	-0,47	0,07	3,1	0,1
contro gli americani	0,08	-0,48	0	0,9
contro l Iraq	-0,31	0,06	2,7	0,1
contro tutte le guerre	0,68	0,63	2,2	2,1
criminale di guerra	-0,59	-0,17	2,6	0,2
dichiarazione di guerra	-0,56	0,62	1,4	2,1
diritto internazionale	-0,41	0,15	1,7	0,3
favorevole alla guerra	0,42	0,15	1,1	0,2
fermare la guerra	0,07	-0,47	0,1	6,3
grande guerra	0,36	0,57	0,6	1,7
guerra breve	0,05	-0,67	0	2,1
guerra dei trent'anni	0,81	-0,17	2,3	0,1
guerra di aggressione	-0,44	-0,28	1,5	0,7
guerra di Bush	0,23	-0,15	0,7	0,4
guerra di difesa	0,49	0,89	0,7	2,8
guerra di liberazione	0,22	0,14	0,8	0,3
guerra è guerra	0,62	-0,27	1,3	0,3
guerra finisca	0,52	-0,26	1	0,3
guerra giusta	0,52	1,01	4,4	19,5
guerra illegale	-0,43	0,02	1	0
guerra illegittima	-0,1	-0,41	0,1	1,1
guerra in Afghanistan	-0,39	0,18	1	0,3
guerra irachena	-0,07	-0,56	0	3,2
guerra legale	0,94	-0,16	3	0,1
guerra lunga	0,39	-0,45	1,7	2,8
guerra nucleare	-0,51	-0,01	1	0
guerra santa	0,26	0,17	0,8	0,4
guerra voluta	0,18	-0,42	0,1	0,8
guerre giuste	0,43	0,35	0,9	0,7
guerre mondiali	-0,15	-0,51	0,1	1,4
in caso di guerra	-0,63	-0,07	1,5	0
legittima difesa	0,78	0,06	2,2	0
liberazione dell Iraq	0,68	0,09	1,6	0
movimento contro la guerra	-0,78	0,12	3,1	0,1
Nazioni Unite	-0,12	0,35	0,2	1,6
no alla guerra	0,25	-0,02	2,5	0
nostro paese	0,4	0,59	1,1	2,8
nuova guerra	-0,48	0,07	1,6	0
opposizione alla guerra	-0,31	-0,25	0,8	0,6
popolazione civile	0,42	0,96	0,7	4,3
popolo iracheno	0,38	-0,19	2,1	0,6
prigionieri di guerra	-0,31	0,18	1,6	0,6
prossima guerra	-0,58	-0,02	1,1	0
quarta guerra mondiale	-0,13	-0,31	0,1	0,8
ripudia la guerra	-0,03	-0,35	0	1,8
scoppiata la guerra	0,47	-0,48	0,7	0,9
seconda guerra del golfo	-0,28	-0,54	0,3	1,3
termine guerra	0,57	-0,2	1,1	0,2
Unione Europea	-0,25	0,05	0,7	0
uranio impoverito	-0,61	-0,15	1,7	0,1
vittime civili	0,57	-0,13	2,8	0,2

Tavola 7. Analisi delle corrispondenze dei segmenti ripetuti a soglia 20 per newsgroups: coordinate e contributi assoluti dei segmenti


Graf. 1 - Proiezione dei segmenti sul piano fattoriale degli assi 1 e
F2 (19,17 %)
F1 (22,37%)
guerra_giusta
popolazione_civile
guerra_di_difesa
CATTOLICI
contro_tutte_le_guerre
nostro_paese
grande_guerra
guerre_giuste
guerra_santa_
favorevole_alla_guerra
concetto_di_guerra
liberazione_dell_Iraq
legittima_difesa
guerra_di_liberazione
POLONIA
vittime_civili_
guerra_dei_trent_anni
guerra_di_Bush
termine_guerra
guerra_legale
LEGA-NORD
popolo_iracheno
guerra_finisca
guerra_è_guerra
guerra_voluta
guerra_lunga_
fermare_la_guerra
contro_gli_americani
scoppiata_la_guerra
guerra_breve
dichiarazione_di_guerra
Amnesty_International
Nazioni_Unite
prigionieri_di_guerra
diritto_internazionale
guerra_in_Afghanistan
Consiglio_di_Sicurezza
INTERNAZIONALE
amministrazione_Bush
contrario_alla_guerra
movimento_contro_la_guerra
Unione_Europea
contro_l_Iraq
guerra_nucleare
guerra_illegale
DESTRA
prossima_guerra
in_caso_di_guerra
uranio_impoverito
criminale_di_guerra
RIFONDAZIONE
opposizione_alla_guerra
PDS
quarta_guerra_mondiale
ripudia_la_guerra
guerra_di_aggressione
a_guerra_finita
guerra_illegittima
guerre_mondiali
seconda_guerra_del_golfo
guerra_irachena
1,2
1
0,8
0,6
0,4
0,2
0
-0,2
-0,4
-0,6
-0,8
-1
1,2
1
0,8
0,6
0,4
0,2
0
-0,2
-0,4
-0,6
-0,8

I contenuti emergono più chiaramente esaminando la proiezione dei segmenti ripetuti sul piano fattoriale formato dagli assi 1 e 2 (Graf. 1). Nel primo quadrante (++), sul quale è collocato il newsgroup it.politica.cattolici troviamo segmenti che rappresentano posizioni diversificate intorno al tema "umanitario" della guerra: <guerra giusta>, <contro tutte le guerre>, <popolazione civile>, <guerra santa>, <concetto di guerra>, <liberazione dell'Iraq>.

Nel quadrante opposto (--) troviamo invece temi di più marcata connotazione politica che esprimono una opposizione netta a questa <seconda guerra del golfo>: <guerra illegale>, <guerra illegittima>, <guerra di aggressione>. In questo quadrante si collocano i due newsgroups che fanno riferimento alla opposizione di sinistra: it.politica.pds e it.politica. rifondazione.

Sul secondo quadrante (+ -) formato dal semiasse positivo dell'asse 1 e dal semiasse negativo dell'asse 2, troviamo segmenti che rievocano le motivazioni della guerra: <guerra di Bush>, <guerra voluta>, <contro gli americani>, <popolo iracheno> <vittime civili>; non mancano alcune riferimenti che sembrano giustificare l'intervento militare: <guerra legale>, <guerra è guerra>. Anche qui si individuano dei segmenti che denotano un dibattito tra favorevoli e contrari alla guerra: <fermare la guerra>, <guerra finisca>, <no alla guerra>, <guerra breve>, <guerra lunga>.

Sul quarto quadrante (- +) , formato dal semiasse negativo dell'asse 1 e dal semiasse positivo dell'asse 2, troviamo segmenti che fanno riferimento agli organismi internazionali (<Nazioni Unite>, <Consiglio di sicurezza>, <Unione Europea>, <Amnesty International>) e alla collocazione della guerra nel contesto della legalità internazionale.

4. Analisi delle corrispondenze e classificazione delle concordanze.

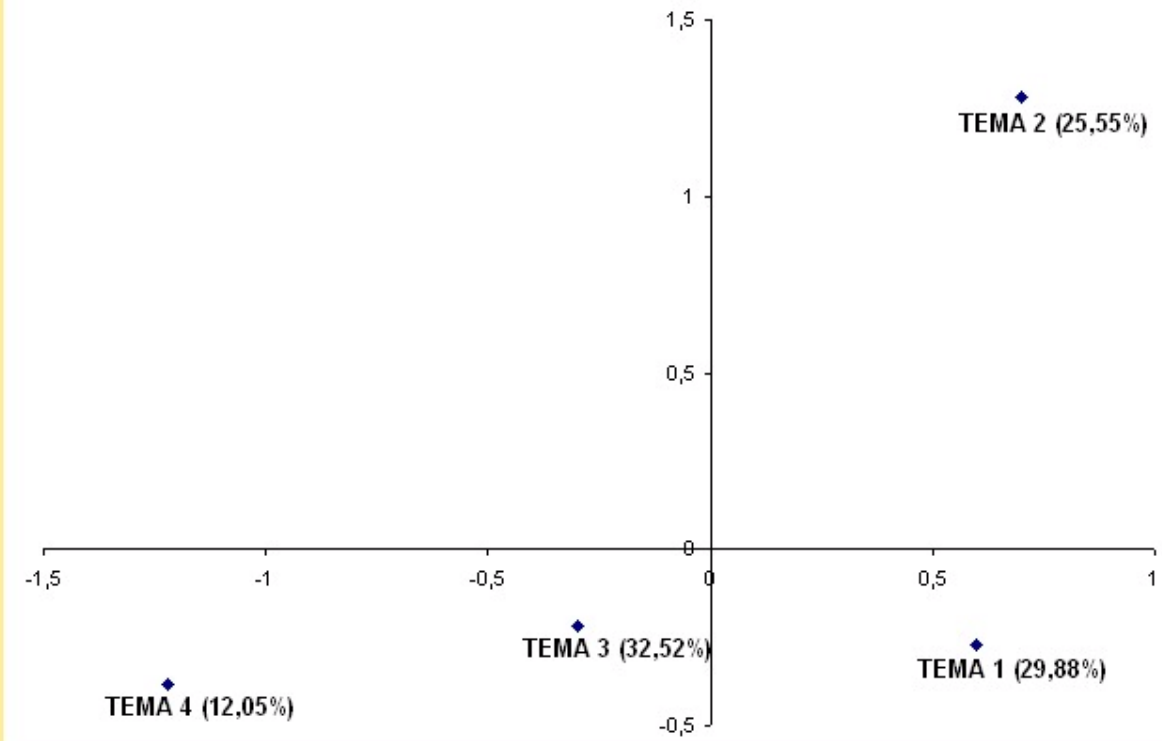
L'individuazione dei temi dibattuti nei newsgroups può essere effettuata anche attraverso l'analisi delle co-occorrenze all'interno delle concordanze della forma <guerra>. L'analisi delle corrispondenze sulla tabella lessicale delle forme per concordanze (mots-réponses) effettuata alla soglia di frequenza 50 con la fase ASPAR del programma SPAD-T 1.5 per DOS ha permesso di individuare due fattori principali che, anche in questo caso, non permettono di evidenziare temi caratteristici di ciascun newsgroup ma permettono di individuare con maggiore dettaglio i nuclei di significato che sono stati dibattuti nel periodo preso in esame.

Per una sintesi più efficace di questo risultato si è ritenuto opportuno procedere ad una classificazione delle concordanze. L'ipotesi-guida di questa strategia di analisi è che dalla individuazione dei temi trattati sia possibile risalire ad una quantificazione, sebbene molto approssimata, del peso di ciascuno di essi nei messaggi del periodo preso in esame. Il peso dei gruppi tematici è rappresentato dalla frequenza relativa delle concordanze classificate all'interno dei rispettivi gruppi.

Nella *cluster analysis* effettuata sulle concordanze sono stati individuati 4 gruppi tematici.

Si riportano qui di seguito le descrizioni dei gruppi tematici sul primo piano fattoriale. Le forme caratteristiche che descrivono ciascun gruppo di concordanze sono elencate in ordine di specificità con valori del test sempre inferiori a 0,01. Le specificità sono state ottenute con la seguente catena di procedure: RECIP – PARTI – MOTEX – MOCAR.

Graf. 2 - Proiezione dei gruppi tematici sul piano fattoriale degli assi 1 e 2



Gruppo tematico 1. Riferimenti alla politica estera italiana e alla polemica tra maggioranza e opposizione (29,88% delle concordanze)

Onu, giusta, Papa, esistere, sinistra, Unione_Europea, posizione, Francia, risoluzione, dire, Italia, opposizione, Germania, preventiva, partecipare, Dio, legittima, dottrina, internazionale, morale, giuste, parole, favorevole, politico, dire_che, legale, politica, dovere, D'Alema, contrari, parola, nome, Fassino, illegale, chiesa, Berlusconi, impedire, porre, potere, costituzione, problema, unilaterale, Cofferati, concetto, umanità, umanitario, illegittimo, pace.

Gruppo tematico 2. Riferimenti alla polemica sul pacifismo e sulle sue motivazioni (25,55% delle concordanze)

in piazza, contro, scendere, manifestazione, popolo, pace, liberazione, manifestare, Saddam, Iraq, credere, andare, Paese, Roma, bandiera, pacifisti, protesta, persona, post, corteo, questa, mondo, motivazione, pacifista, obiettivo, aprile, usare, gente, unità, sicuro, iracheno, ragione, a_favore_di, giovani, governo, intendere, prova, opinione, scrivere, continuare, liberare, campagna, dittatore, volere, santa, nazionale, esempio, portare, appoggiare, discutere, lotta.

Gruppo tematico 3. Riferimenti alla politica internazionale americana e alle motivazioni dell'intervento militare (32,52% delle concordanze)

mondiale, prima_di, seconda_guerra_mondiale, perdere, guerra_fredda, crimine, militare, americano, alleati, vincere, quarta, Stati_Uniti, durare, prima_guerra_mondiale, uomo, nemici, venire, sangue, anno, potenza, alla_fine_di, effetto, ogni, sperare, terrore, terza, amministrazione, scoppiare, storia, mese, secolo, Bush, terribile, stare, passare, minaccia, tre, statunitense, petrolio, politici, due, nemico, paura, criminale, rispondere, sapere, fronte, Siria.

Gruppo tematico 4. Riferimenti alle cronache della guerra e al regime di Saddam (12,05% delle concordanze)

civili, morti, di_massa, guerra_del_golfo, vittima, distruzione, Afghanistan, armi, durante_la_guerra, bambino, città, pentagono, soldato, bomba, bombardamento, centinaio, tornare, iracheni, missili, militari, Iran, notizia, golfo, milione, migliaio, decina, occidentali, danni, colpire, Bagdad, arrivare, propaganda, popolazione, giorno, donna, giornalista, prigioniero, morire, morte, truppa, Vietnam, vita, capo, tv, affare, embargo, raccontare, sporca, totale.

Il tema prevalente è il n. 3 (32,52% delle concordanze). Le forme grafiche classificate in questo gruppo individuano l'argomento che ha interessato la maggior parte degli utenti dei newsgroups: il dibattito intorno alle motivazioni dell'intervento militare e alle sue giustificazioni sul piano del diritto internazionale. Emergono diversi riferimenti ai crimini compiuti da Saddam Hussein, alle scelte dell'amministrazione Bush, alla storia delle guerre mondiali che hanno caratterizzato il XX secolo, allo "scontro tra civiltà". Le motivazioni vengono ricondotte alla paura del terrorismo, alla liberazione dell'Iraq dalla dittatura, quanto alla presenza dei pozzi petroliferi.

5. Conclusioni

La strategia di analisi adottata si è rivelata adeguata alla individuazione dei temi principali del dibattito nei newsgroups presi in esame. La presenza di alte componenti di "rumore" nei messaggi è stata neutralizzata con l'analisi delle concordanze della forma <guerra> senza comportare perdite significative per l'analisi del contenuto. Le scelte compiute con l'utilizzazione di TALTAC come strumento di riconoscimento delle forme grafiche, normalizzazione, disambiguazione e tagging ha dato dei buoni risultati anche sul piano della "pulitura" automatica del testo.

Nei messaggi non è stato possibile individuare tendenze ideologico-politiche distinte in sintonia con le scelte tematiche dei newsgroups. Dall'analisi delle concordanze non è emerso un lessico specifico della guerra in relazione con le presunte appartenenze ideologiche degli utenti quanto, piuttosto, in relazione con i temi individuati che sono trasversali rispetto ai newsgroups stessi. Gli utenti amano intrattenersi in un dibattito politico attento, informato, vivace e caratterizzato da un elevato tono polemico. Tuttavia non emergono comportamenti verbali significativamente oltraggiosi se non in misura del tutto occasionale e frutto di provocazioni estranee al clima generale. Saranno necessarie ulteriori analisi sul corpus complessivo dei newsgroups, e non soltanto sul sub-corpus delle concordanze, per poter confermare o meno l'esistenza di un lessico della guerra caratteristico delle diverse posizioni ideologiche.

Bibliografia

Beaudouin V., Fleury S. e Velkowska J. (2000). Études des échanges électroniques sur internet et intranet : forums et couriers électroniques. In *Actes des JADT 2000*, vol. (1) : 17-24.

Bolasco S. (1999). *Analisi multidimensionale dei dati*. Carocci.

Bolasco S., Baiocchi F. e Morrone A. (2000-2003). *TALTAC 1.6. Trattamento Automatico Lessico Testuale del Contenuto*. CISU.

Choi J.H. e Danowski J. (2003). Making a Global Community on the Net – Global Village or Global Metropolis ? : A Network Analysis of Usenet Newsgroups. *Journal of Computer Mediated Communication*, vol. (7/3). http://www.ascusc.org/jcmc/vol7/issue3/choi.html.

Giuliano L. (2002). G8-2001 : la rivolta nel monitor. Analisi testuale dei messaggi nel newsgroup <it.eventi.g8.genova> durante gli scontri di piazza. In *Actes des JADT 2002* : 301-311.

Lebart L., Morineau A., Becue M. e Hausler L. (1993). *SPAD-T 1.5 (DOS*). CISIA.

Lamalle C., Martinez W., Fleury S. e Salem A. (2002). *Lexico3.* Université de la Sorbonne Nouvelle – Paris 3.

Rosen D., Woelfel J., Krikorian D. e Barnett G.A. (2003). Procedures for Analyses of Online Communities. *Journal of Computer Mediated Communication*, vol. (8/4). http://www.ascusc.org/jcmc/vol8/issue4/rosen.html.

Smith M.A. (1999). Invisible crowds in cyberspace : Mapping the social structure of the Usenet. In Smith M.A. e Kollock P. (Eds), *Communities in cyberspace*. Routledge : 195-219.

Smith M.A. (2002). Mapping Social Cybercpaces : Measures and Maps of Usenet, a Computer Mediated Space. *Dissertation Abstracts International, A : The Humanities and Social Sciences*, vol. (63/1) : 382-A.

Stubbs P. (1998). Conflict and Co-Operation in the Virtual Community : eMail ed the Wars of the Yugoslav Succession. *Sociological Research Online*, vol. (3). http://www.soc.surrey.ac.uk/socresoline/.

Tuzi A. (2003). *L'analisi del contenuto*. Carocci.

Whittaker S., Terveen L., Hill W. e Chemy L. (1998). The dynamics of mass interaction. In *Proceedings of Conference on Computer Supported Cooperative Work* : 257-264. http://www.acm.org/pubs/citations/proceedings/ cscw/289444/p257-whittaker/.

Generative vs Discriminative Approaches to Entity Recognition from Label-Deficient Data

Cyril Goutte, Éric Gaussier, Nicola Cancedda, Hervé Déjean

Xerox Research Centre Europe – 6, ch. de Maupertuis – 38240 Meylan – France
{Cyril.Goutte,Eric.Gaussier,Nicola.Cancedda,Herve.Dejean}@xrce.xerox.com

Abstract

Annotating biomedical text for Named Entity Recognition (NER) is usually a tedious and expensive process, while unannotated data is freely available in large quantities. It therefore seems relevant to address biomedical NER using Machine Learning techniques that learn from a combination of labelled and unlabelled data. We consider two approaches: one is discriminative, using Support Vector Machines, the other generative, using mixture models. We compare the two on a biomedical NER task with various levels of annotation, and different similarity measures. We also investigate the use of Fisher kernels as a way to leverage the strength of both approaches. Overall the discriminative approach using standard similarity measures seems to out-perform both the generative approach and the Fisher kernels.

Résumé

L'annotation de textes médicaux pour la reconnaissance d'entités nommées est un travail pénible et couteux, alors que, au contraire, de larges quantités de données non annotées sont disponibles gratuitement sur le Web, par exemple sur MedLine/PubMed. Il est donc particulièrement pertinent de s'attaquer au problème de la reconnaissance d'entités biomédicales à l'aide de techniques d'apprentissage automatique qui apprennent à partir de mélanges de données annotées et non annotées. Nous considérons deux approches: l'une est discriminative, elle repose sur l'utilisation de modèles à point support ou SVM ; l'autre est générative et utilise des modèles de mélange. Nous comparons les deux approches sur une tache de reconnaissance d'entités biomédicales avec divers niveaux d'annotation, et en utilisant différentes mesures de similarité. Nous étudions aussi l'utilisation des noyaux de Fisher afin de tenter de combiner les atouts de chacune des deux approches. Dans l'ensemble, l'approche discriminative utilisant des mesures de similarité standard semble plus performante que l'approche génerative, ainsi que l'utilisation des noyaux de Fisher.

Keywords: biological entity recognition, partially labelled data, discriminative, generative, support vector machines, transductive inference, mixture models, fisher kernels.

1. Introduction

Entity recognition is a crucial step in information extraction. In technical domains such as biomedicine, it is often necessary to recognise specific entities such as protein or gene names. Machine Learning techniques are therefore attractive, as they allow to automatically learn an entity recognition engine for a new domain, with minimal involvement from the user.

One often cited drawback of the Machine Learning approach is that it relies on a corpus, which is usually annotated manually, often a tedious and costly task, especially in technical domains. On the other hand, unannotated data is usually relatively plentiful and freely available, eg in the biomedical domain (cf. http://www.pubmed.org).

As a consequence, it seems particularly relevant to investigate the use of recent Machine Learning approaches that learn from a combination of labelled and unlabelled data, to tackle the

Named Entity Recognition (NER) problem in technical domains. Such approaches include transductive inference (Vapnik, 1995) or learning probabilistic models from partially labelled data (Miller and Uyar, 1997). The former is used in a discriminative setting, estimating directly the labelling of the unlabelled data, while the latter is generative, estimating a probabilistic model that can generate both labelled and unlabelled data. As a dataset containing both labelled and unlabelled data may be seen as a dataset with "missing" labels, it is also called *label-deficient*.

In this contribution, we wish to explore the differences, and the possible links, between these discriminative and generative approaches, in the context of NER. Section 2 introduces the NER problem in the framework of (supervised) categorisation. Section 3 briefly reviews the discriminative and generative approaches, and describes several techniques: transductive inference for Support Vector Machines (3.1.), probabilistic models of label-deficient data (3.2) and Fisher kernels (3.3). In section 4, we explore the experimental results obtained with these approaches on a biological entity recognition task, and we conclude with a discussion.

2. Categorisation for Named Entity Recognition

In our work, we focus on the problem of recognising specific biomedical entities from abstracts of scientific articles. For example, in:

> Inhibition of the activity of Drosophila suppressor of *hairless* and of its human homolog, *KBF2/RBP-J kappa*, by protein interaction.

we wish to recognise that *hairless* and *KBF2/RBP-J kappa* are gene names, while all other terms are irrelevant. We focus on identifying names of genes, proteins or RNA, all other terms (species, chemical names) being irrelevant for this task. Note, however, that we adopt a general Machine Learning approach, which should be applicable to other classes of entities, provided the necessary (small) amount of annotation is available. In addition, we are more interested in the comparison between generative and discriminative approaches using label-deficient data than to the raw performance of either approaches. Several other factors, such as the choice of the appropriate feature set, may influence the final performance of the system.

Many successful approaches to NER formulate the problem in terms of categorisation: is a candidate term an entity (category 1, relevant) or not (category 0, irrelevant)? Examples of this include most of the contributions to the shared task of the last two CoNLL conferences (Roth and van den Bosch, 2002; Daelemans and Osborne, 2003) and several approaches to Biomedical NER (Kazama *et al.*, 2002; Lee *et al.*, 2003; Takeuchi and Collier, 2003; Yamamoto *et al.*, 2003).

In our case, gene, protein and RNA names are relevant, all other terms are not. Given an example $x \in \mathcal{X}$, the entity recognition problem is formulated as the problem of learning a categoriser $h : \mathcal{X} \to \{0; 1\}$, such that the probability that a term is recognised correctly, $P(h(x) = y)$, is maximised over the distribution of (x, y).

3. Learning from partially labelled data

In many text processing applications, it is easy to obtain large amounts of unannotated data, and it may be costly to annotate them. Machine Learning techniques that learn from partially labelled data have therefore been applied to eg text categorisation with some success (Joachims, 1999; Nigam *et al.*, 2000).

The usual way to learn the function h that recognises the relevant entitites is through a sample

of annotated data $(x^{(i)}, y^{(i)})$. It had been argued that having large amounts of non annotated data in the form of additional data $(x^{(j)})$ could in principle help better characterise the classes and therefore improve the categorisation abilities. Imagine for example that the data is formed of a small number of well separated components, each corresponding to a category. The components can be modelled arbitrarily well using unlabelled data alone (possibly in large quantities). Once the components are modelled, little labelled data would suffice to assign a proper label to each of them. If the components are well approximated, the performance would be good, much better than when the classes are learned on the basis of labelled data alone.

This justification is generative in nature, as it relies on the ability to model the data and its labelling. In standard classification problems, however, it has been noted that discriminative methods often out-perform generative techniques, especially in the limit of large sample sizes (Rubinstein and Hastie, 1997; Ng and Jordan, 2002). With partially labelled data, there is also a distinction between discriminative and generative methods. Support Vector Machines are discriminative, and may handle unlabelled data using transductive inference (sec. 3.1.), while probabilistic models of label-deficient data are generative (sec. 3.2.). Fisher kernels (sec. 3.3.) represent a way to bridge the two, by using the generative model to derive a similarity that is used in the discriminative approach.

3.1. *Transductive inference for discriminant analysis*

The standard statistical learning paradigm is to induce a model (eg a Support Vector Machine) from the data and deduce the labelling of test data from this model. Vapnik (1995) argues that the model may actually be superfluous, and advocates the use of *transductive learning* to directly estimate the labelling without first learning the model from the training data. Given an input test data x^*, the transductive approach (Vapnik, 1995; Saunders *et al.*, 1999) tries to estimate directly the decision y^* rather than induce a model $\widehat{h}$ and infer that $y^* = \widehat{h}(x^*)$. Whith little unlabelled data, this is relatively straightforward, but quickly becomes impractical as the number of unlabelled examples increases. Fortunately, using a few heuristics, it is possible to provide an efficient approximate solution to transductive inference with SVM (Joachims, 1999b). In our work, we use this solution, as implemented in Thorsten Joachims' SVMlight software (Joachims, 1999a).

The algorithm starts by labelling the unlabelled data using a SVM trained inductively on the labelled data alone, then repeatedly infer a SVM on the data with the completed labels and tries to improve on the solution by swapping the labels of pairs of examples. This algorithm does converge towards a stable, although obviously not necessarily optimal, solution. Using transductive inference, unlabelled data are taken into account as "test data", and although the inferred labels are usually not used, they have an influence both on the labelling of the true test data, and on the resulting model.

3.2. *Generative models for label-deficient data*

Miller and Uyar (1997) proposed a mixture model to handle combinations of labelled and unlabelled data. Both data and labels are assumed to be generated independently by components of the mixture: $P(x, y) = \sum_\alpha P(\alpha)P(x|\alpha)P(y|\alpha)$. By marginalising over labels y, $P(x) = \sum_\alpha P(\alpha)P(x|\alpha)$. In order to model continuous data, Miller and Uyar (1997) used Gaussian mixtures for the data-dependent component distributions $P(x|\alpha)$. Applications to text processing typically rely on mixtures of binomial distributions. Here, we use a co-occurrence model (Hofmann, 1999) in which each observation x is the co-occurrence of an entity e and a feature f, and $P(x|\alpha) = P(e|\alpha)P(f|\alpha)$. Given a dataset $\mathcal{D}$ composed of labelled and unla-

belled data, $\mathcal{D} = \mathcal{L} \cup \mathcal{U}$, with $\mathcal{L} = \left\{(x^{(i)}, y^{(i)})\right\}_{i=1\ldots N}$ and $\mathcal{U} = \left\{x^{(i)}\right\}_{i=N+1\ldots M}$, parameters are estimated by maximising the (log-)likelihood:

$$L(\mathcal{D}) = \sum_{i=1}^{N} \ln P(x^{(i)}, y^{(i)}) + \sum_{i=N+1}^{M} \ln P(x^{(i)})$$

This may be performed using 2 variants of the Expectation-Maximisation algorithm (Dempster *et al.*, 1977), depending on which latent variables are considered: components alone or components and labels. Both variants maximise the same likelihood and therefore yield similar results. Here we use the former, aka EM1 (Miller and Uyar, 1997), see appendix for details.

Enforcing no constraint on the label generation (so-called *soft partitioning*), especially when classes are unbalanced, results in *cluster impurity*: all or most components contain significant portions of examples from all classes. Components are therefore badly aligned with the discrimination task, and the model yields bad categorisation performance. In order to avoid this issue in the naturally unbalanced problem of NER, we use *hard partitioning*: components are tied to a specific class, ie $P(y|\alpha)$ take binary 0/1 values.

Once the model parameters have been estimated, a label may be assigned to a new example x according to $P(y|x) = (\sum_\alpha P(\alpha)P(x|\alpha)P(y|\alpha))/(\sum_\alpha P(\alpha)P(x|\alpha))$.

3.3. *Fisher kernels*

Jaakkola and Haussler (1999) introduced Fisher kernels as a similarity measure derived from a probabilistic model. This may be useful whenever a probabilistic model of the data exists, as the model-induced similarity may be different from the similarity between observed features. Denoting the log-likelihood for example x by $\ell(x) = \ln P(x|\theta)$, and using the Fisher information matrix $\mathbf{I_F} = \mathsf{E}\left(\nabla\ell(x)\nabla\ell(x)^\top\right)$, the Fisher kernel is:

$$FK(x_1, x_2) = \nabla\ell(x_1)^\top \mathbf{I_F}^{-1} \nabla\ell(x_2) \tag{1}$$

For a suitable choice of parameterisation, $\mathbf{I_F}$ is usually approximated by the identity. In the context of our work there are at least two ways to use Fisher kernels: we can either learn an unsupervised model of the input data $P(x)$ alone, or learn a label-deficient mixture model using the partially annotated data as explained above. In both cases, we then derive a Fisher kernel using equation 1. In our case, these two alternatives correspond to PLSA (Hofmann, 2000) and to the EM1 mixture model. It turns out that the parameters in both models are identical, although they are estimated by maximising different likelihoods, and therefore lead to different parameter estimates. The expression for the Fisher kernels, however, is identical:

$$FK(x_1, x_2) = \sum_\alpha \frac{P(\alpha|e_1)P(\alpha|e_2)}{P(\alpha)} + \sum_f \widehat{P}(f|e_1)\widehat{P}(f|e_2) \sum_\alpha \frac{P(\alpha|f, e_1)P(\alpha|f, e_2)}{P(f|\alpha)} \tag{2}$$

$\widehat{P}(f|e_1)$ (resp. $\widehat{P}(f|e_2)$) is the normalised observed frequency of feature f for entity e_1 (resp. e_2) corresponding to example x_1 (resp. x_2), and all other parameters are obtained during training. Although the expression is identical for PLSA and for EM1, the kernels differ through different parameter estimates. In particular, hard partitioning in EM1 ensures that $FK(x_1, x_2) = 0$ for all pairs of examples with differing labels.

Feature	Value
FULL	1
%LexGENE	1
SYNONYM	1
DICOAMB	1

Feature	Value
LEMME_hairless	1
ADJ	1
LC_Drosophila	1
LC_suppressor	1
LC_PREP	2

Feature	Value
RC_CONJ	1
RC_PREP	1
RC_PRON	1
RC_human	1

Table 1. Features of hairless *in* "of Drosophila suppressor of hairless and of its human"

4. Experimental results

Our dataset was formed using 184 abstracts queried from MedLine. These abstracts were manually annotated by a trained biologist with various biomedical entities. Of these, we focus on gene, protein and RNA names only. We use 122 abstracts as development set (training and validation) and 62 abstracts for testing. All abstracts are tokenised, lemmatised and tagged with part-of-speech information using Xerox linguistic tools. In addition we apply a pre-filtering step that discards all terms that are in a dictionary of common English and are not in the dictionary of possible biological terms. The assumption behind this pre-filtering step is that biological terms that are ambiguous with general English words are well identified and therefore will be in our biomedical resources. Indeed, this filter discards 80% of the original tokens (31617 out of 40398) with a recall of 94% on relevant tokens (2394 out of 2550). We end up with 8781 candidates: 5865 in the development set and 2916 in the test set.

For all candidates, we generate four types of features: spelling (uppercase, digits, etc.), lexical (presence in various dictionaries), linguistic (part-of-speech and lemma) and contextual (words in a 4-word context on either side of the candidate). Table 1 shows all the non-zero features for one of the words in our earlier example. In total, there are 7277 features, although few of them are non-zero for each candidate.

In order to analyse the behaviour of various methods in the presence of labelled and unlabelled data, we retain a variable proportion of the annotation, between 1% and 67%. For each level of partial annotation, we sample 10 different sets of labels, and average the performance over these 10 samples.

In all cases, the baseline is a simple dictionary lookup, combined with part-of-speech information: a candidate is tagged as a relevant entity iff it is present in one of the dictionaries of relevant entities *and* it is tagged as a noun. The precision/recall of this baseline are 58.44%/86.10%, giving a F_1 score of 69.62%.

We first consider Support Vector Machines trained with inductive and transductive inference, and the mixture model with hard partitioning, trained with EM1. Figure 1 shows that transductive inference consistently outperforms both inductive inference and the mixture model. Interestingly, the performance of transductive inference is always better than the baseline, except for one point (RBF kernel at 1% annotation). To put these results into prospective, consider that 2% annotation corresponds to 117 annotated examples (31 positives, 86 negatives) or around $2\frac{1}{2}$ abstracts, ie very little actual annotation work.

All methods yield very similar performance above 16% annotation, but below this level, the performance of transductive inference degrades much more gracefully. Inductive inference and the mixture model need 4 to 8% of annotation to outperform the baseline. Note that although the baseline uses no training data *per se*, it actually relies on a fair amount of prior knowledge on the task and on the available ressources. On the other hand, the Machine Learning method used

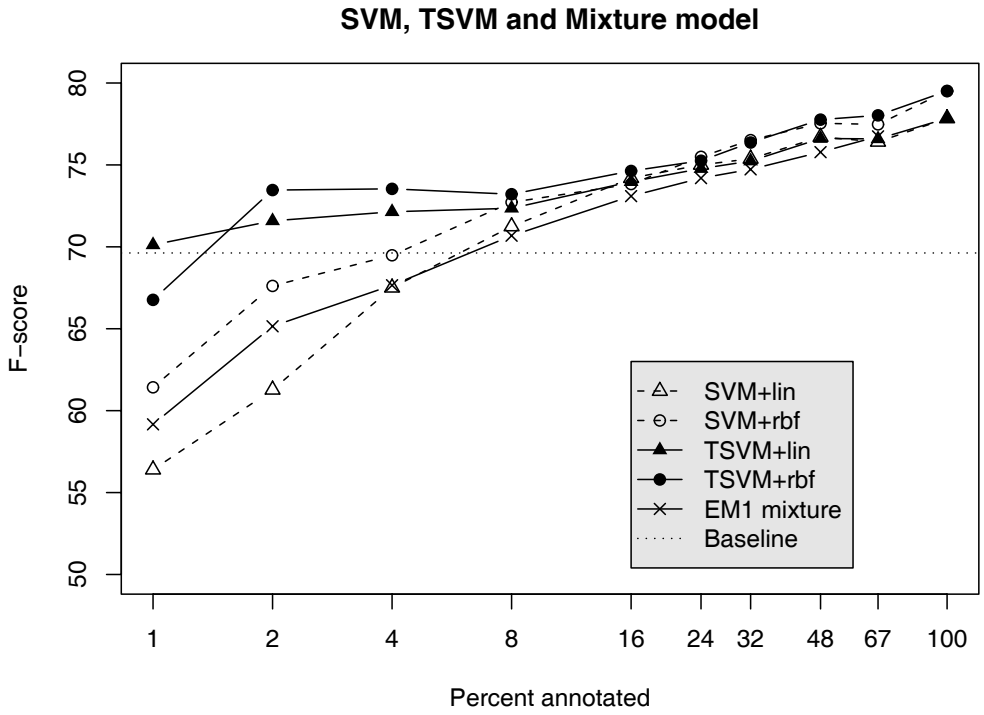


Figure 1. Performance of SVM with lin(ear) and rbf kernels trained using inductive and transductive inference, and mixture model for label-deficient data (EM1). Median F-score over 10 random samples.

here rely only on the available data, and do not use any prior information beside the choice of the feature set. It is therefore not counter-intuitive that the baseline should manage to outperform some of the ML methods when little annotated data is available.

We then address a second relevant question: can we bridge the gap between the discriminative SVM and the generative mixture model using eg Fisher kernels as a similarity measure. In order to investigate this, we calculated the SVM associated with the purely unsupervised PLSA model (FK0) and with the semi-supervised mixture trained by EM1 (FK1). Both kernels are used to train SVM with both inductive and transductive inference. The results are displayed in figure 2. For comparison, we plot the results obtained with the RBF kernel, the best performing of the standard kernels in figure 1.

Clearly the performance of the Fisher kernel derived from PLSA is dire, although transductive inference seems to help. We attribute the poor performance to the problem of cluster impurity, which prevents the model from obtaining components which really correspond to identifiable labels. The performance of the Fisher kernel trained on the label-deficient mixture (FK1) is close to the performance of the RBF kernel, although it is consistently inferior. Finally, FK1 is the only kernel for which transductive inference does not increase, but rather decreases performance.

5. Discussion and conclusion

In this paper, we investigated the use of discriminative versus generative techniques for learning NER from partially labelled data. Based on our experimental results, the discriminative SVM seem to perform significantly and consistently better than the generative mixture model. In fact the mixture model, even though it uses additional unlabelled data, does not manage to

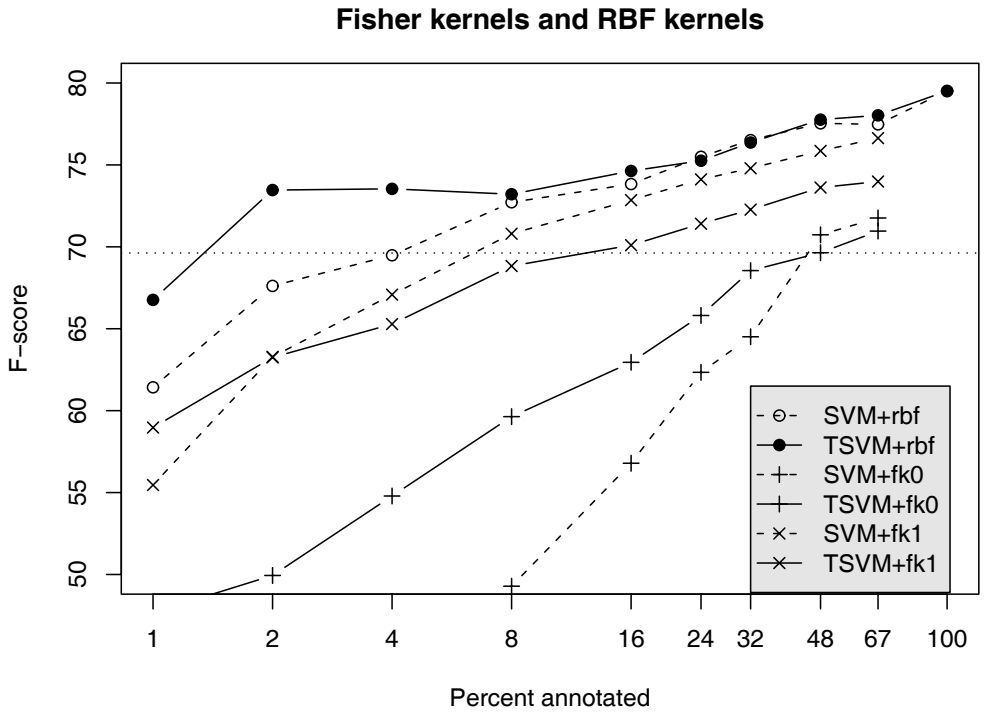


Figure 2. Comparison between Fisher kernels derived from PLSA (FK0), from the semi-supervised mixture (FK1), and standard RBF kernel. Median over 10 random samples.

outperform the SVM trained by inductive inference on the labelled data only. The discriminative approach using transductive inference manages to outperform the baseline using as little as 1% of annotation, ie slightly more than 1 abstract, while all other approaches need at least 4 to 8 times as much data to reach that level of performance. Transductive inference is consistently at least as good, and often significantly better, than inductive inference, for almost all kernels, with the notable exception of the Fisher kernel derived from the generative model.

Overall, the standard RBF kernel performs best. This may be attributed to the fact that it performs an implicit feature expansion into an infinite dimensional space, while all other kernels that we tried (linear and Fisher kernel) use a low-dimensional feature space. It may therefore be easier to "pick" efficient dimensions from the infinite number of implicit RBF dimension, than to work with the relatively few available dimensions in all other cases. The fisher kernels never manage to outperform the standard kernel in our experiments. For FK0, this is mostly due to the *cluster impurity* problem. For FK1, we believe that this is due to the geometry of the implicit feature space when the model uses hard partitioning. In that case, the annotated positive and negative examples belong to two simplexes in two orthogonal subspaces, such that the decision function outside the space spanned by the annotated examples is essentially arbitrary.

These results suggest that using partially labelled data efficiently may yield large performance gains. To illustrate this further, future experiments will apply these techniques to the full annotated dataset, using additional unannotated data queried from PubMed.

Acknowledgements

We thank Anne Schiller, Ágnes Sandor and Violaine Pillet for help with the data. This research was supported by the EC under the KerMIT project (IST-2001-25431).

Appendix: EM equations for label-deficient data

The EM equations for EM1 and hard partitioning are:

$$C^{(t)}(\alpha,i) = \langle P(\alpha|x^{(i)},y^{(i)})\rangle = \frac{P(\alpha)P(x^{(i)}|\alpha)P(y^{(i)}|\alpha)}{\sum_\alpha P(\alpha)P(x^{(i)}|\alpha)P(y^{(i)}|\alpha)} \qquad \text{E-step, labelled data} \qquad (3)$$

$$C^{(t)}(\alpha,i) = \langle P(\alpha|x^{(i)})\rangle = \frac{P(\alpha)P(x^{(i)}|\alpha)}{\sum_\alpha P(\alpha)P(x^{(i)}|\alpha)} \qquad \text{E-step, unlabelled data} \qquad (4)$$

$$P^{(t+1)}(\alpha) = \frac{1}{N}\left(\sum_i C^{(t)}(\alpha,i)\right) \qquad \text{M-step for } P(\alpha) \qquad (5)$$

where $\langle\cdot\rangle$ indicates the expectation conditioned on the data and current parameter estimates. The M-step equations for the co-occurrence model $P(x|\alpha) = P(e|\alpha)P(f|\alpha)$ (Hofmann, 1999) are:

$$P^{(t+1)}(e|\alpha) = \frac{\sum_{i,e^{(i)}=e} C^{(t)}(\alpha,i)}{\sum_i C^{(t)}(\alpha,i)} \text{ and } P^{(t+1)}(f|\alpha) = \frac{\sum_{i,f^{(i)}=f} C^{(t)}(\alpha,i)}{\sum_i C^{(t)}(\alpha,i)} \qquad (6)$$

Parameters are obtained by iterating equations 3 through 6, using a deterministic annealing scheme (Ueda and Nakano, 1995) in order to reduce the sensitivity to initial conditions.

References

Ananiadou S. and Tsujii J. (Eds) (2003). In *Proceedings of the ACL workshop on Natural Language Processing in Biomedicine*.

Daelemans W. and Osborne M. (Eds) (2003).In *Proceedings of the Seventh Conf. on Natural Language Learning*.

Dempster A.P., Laird N.M. and Rubin D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, vol. (1): 1-38.

Hofmann T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conf. on Uncertainty in Artificial Intelligence*: 289-296.

Hofmann T. (2000). Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In *Advances in Neural Information Processing Systems*, vol. (12).

Jaakkola T. S. and Haussler D. (1999). Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems*, vol. (11): 487-493.

Joachims T. (1999a). Making large-scale SVM learning practical. In Schölkopf B., Burges C. and Smola A. (Eds), *Advances in Kernel Methods - Support Vector Learning*. MIT-Press.

Joachims T. (1999b). Transductive inference for text classification using support vector machine. In Bratko I. and Dzeroski S. (Eds), *Machine Learning – Proceedings of the 16th Intl Conf.*: 200-209.

Kazama J., Makino T., Ohta Y. and Tsujii J. (2002). Tuning support vector machines for biomedical named entity recognition. In Johnson S. (Ed.), *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain*: 1-8.

Lee K.-J., Hwang Y.-S. and Rim H.-C. (2003). Two-phased biomedical NE recognition based on SVMs. In Ananiadou S. and Tsujii J. (Eds), *Proceedings of the ACL workshop on Natural Language Processing in Biomedicine*: 33-40.

Miller D.J. and Uyar H.S. (1997). A mixture of experts classifier with learning based on both labelled and unlabelled data. In *Advances in Neural Information Processing Systems (NIPS*9)*: 571-577.

Ng A.Y. and Jordan M. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems*, vol. (14).

Nigam K., McCallum A., Thrun S. and Mitchell T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, vol. (39/2-3): 103-134.

Roth D. and van den Bosch A. (Eds) (2002). In *Proceedings of the Sixth Conf. on Natural Language Learning*.

Rubinstein Y. D. and Hastie T. (1997). Discriminative vs informative learning. In *Proceedings of the 3rd Intl Conf. on Knowledge Discovery and Data Mining*: 49-53.

Saunders C., Gammerman A. and Vovk V. (1999). Transduction with confidence and credibility. In Dean T. (Ed.), *Proceedings of the Sixteenth Intl Joint Conf. on Artificial Intelligence*: 722-726.

Takeuchi K. and Collier N. (2003). Bio-medical entity extraction using support vector machines. In Ananiadou S. and Tsujii J. (Eds), *Proceedings of the ACL workshop on Natural Language Processing in Biomedicine*: 57-64.

Ueda N. and Nakano R. (1995). Deterministic annealing variant of the EM algorithm. In *Advances in Neural Information Processing Systems*, vol. (7): 545-552.

Vapnik V.N. (1995). *The Nature of Statistical Learning Theory*. Springer.

Yamamoto K., Kudo T., Konagaya A. and Matsumoto Y. (2003). Protein name tagging for biomedical annotation in text. In Ananiadou S. and Tsujii J. (Eds), *Proceedings of the ACL workshop on Natural Language Processing in Biomedicine*: 65-72.

Relazioni non Simmetriche tra Corpora

Maria Gabriella Grassia, Michelangelo Misuraca, Germana Scepi

Dipartimento di Matematica e Statistica – Università Federico II – Napoli – Italia
michelangelo.misuraca@unina.it

Abstract

In this paper the language used by firms for searching new employers by web is studied. Particularly, we are interesting in evaluating the dependence between two *corpora*, e.g. one defined by the forms used for describing the skills of the candidates for jobs and the other defined by the forms used by firms for describing their mission. The method used is textual data analysis, more precisely, a non symmetrical correspondence analysis on a peculiar lexical table forms/forms with an *ad hoc* weighting system on the explanatory variables. Furthermore, the main results of an application on a sample of firms are showed in terms of friendly readable graphical representations.

Riassunto

In questo lavoro si presenta uno studio sul linguaggio utilizzato dalle aziende alla ricerca di candidati da assumere per differenti mansioni. L'obiettivo è quello di valutare la dipendenza tra due *corpora*, con riferimento particolare alla relazione tra le forme usate per definire le caratteristiche dei candidati richiesti e le forme utilizzate dalle stesse aziende per descrivere la propria *mission* aziendale. Il metodo considerato è l'analisi dei dati testuali e, precisamente, l'analisi non simmetrica delle corrispondenze applicata però, ad una particolare tabella lessicale del tipo forme/forme, con l'introduzione, inoltre, di un sistema di pesi *ad hoc* sulle variabili esplicative. Nel lavoro vengono presentati i principali risultati, in termini soprattutto di rappresentazioni grafiche, conseguiti dall'applicazione su un campione di 167 aziende di diversa dimensione e settore di attività, distribuite su tutto il territorio nazionale.

Keywords: textual data analysis, non symmetrical correspondence analysis, term frequency.

1. Introduzione

Nel presente lavoro l'attenzione è rivolta allo studio del linguaggio utilizzato dalle aziende per assumere nuovo personale. In particolare, si vuole analizzare ed esplicitare la possibile relazione tra il linguaggio che ciascuna azienda usa per descrivere la propria *mission* (che da ora in poi battezzeremo con l'allocuzione "chi siamo") e quello utilizzato per descrivere i profili dei candidati all'assunzione (che battezzeremo con "chi cerchiamo"). L'ipotesi di partenza è che tale relazione non sia di tipo simmetrico bensì che si possa ipotizzare la dipendenza di un linguaggio dall'altro, ossia che il "chi siamo" influenzi il linguaggio utilizzato per descrivere il "chi cerchiamo".

Lo strumento prescelto è l'analisi dei dati testuali vista come estensione di metodi statistici proposti in origine per l'analisi delle relazioni tra variabili numeriche. A differenza della classica matrice di contingenza, del tipo documenti/forme, generalmente analizzata da tale tecnica, nel presente lavoro si definisce una matrice del tipo forme/forme. In particolare, partendo dai due differenti *corpora* rilevati sulle stesse unità statistiche, si costruisce una matrice di *co-presenza* che ha come termine generico il numero di volte in cui le forme dei due *corpora* si presentano contemporaneamente.

L'ipotesi di partenza ci induce ad analizzare tale matrice con una tecnica di analisi dei dati di tipo non simmetrico, e, per la precisione, con l'Analisi non Simmetrica delle Corrispondenze (ANSC, Lauro e D'Ambra, 1984). L'ANSC per tabelle lessicali è stata proposta da Balbi (1995) in alternativa all'uso dell'Analisi delle Corrispondenze (Lebart *et al.*, 1991), laddove le variabili interessate non sembrano essere in una relazione di tipo simmetrico. L'ANSC, inoltre, essendo basata su una metrica euclidea non ponderata, risulta particolarmente adatta per l'analisi di tabelle lessicali ricche di zeri, dove, invece, l'utilizzo di una metrica χ^2 finisce con l'attribuire un'importanza eccessiva alle modalità rare.

Rispetto al metodo originario, nel lavoro si introduce un sistema aggiuntivo di pesi sulle forme del vocabolario del *corpus* del "chi siamo", supposte esplicative rispetto alle forme del vocabolario del "chi cerchiamo". Tale sistema consente di introdurre ulteriori informazioni sulla frequenza di utilizzo delle forme e di migliorare i risultati dell'analisi, sia in termini di leggibilità delle rappresentazioni grafiche che di aumento dell'indice di predittività (τ di Goodman e Kruskal, 1954).

Nel paragrafo successivo (paragrafo 2) verranno introdotti i principi generali dell'ANSC, fornendo regole per l'interpretazione delle rappresentazioni grafiche da essa ottenute. L'interesse e la facilità di lettura di tali rappresentazioni le rendono decisamente uno dei motivi principali dell'utilizzo di tale tecnica nell'ambito dei dati testuali. Nel paragrafo 3 si presenterà la struttura dei dati analizzata e la strategia di analisi adottata. Infine, nel paragrafo 4 verranno mostrati i principali risultati ottenuti dall'applicazione della strategia ad un campione di 167 aziende.

2. Alcuni richiami all'Analisi Non Simmetrica delle Corrispondenze

Consideriamo due variabili qualitative, z_i (con i=1,...,*I*) ed y_j (con j=1,...,*J*), osservate sulle stesse *n* unità e aventi rispettivamente *I* e *J* categorie di risposta. Si classifichino le variabili nella matrice **F**(*I*,*J*) che ha come elemento generico la frequenza relativa congiunta f_{ij}.

Si indichi con **r** il vettore contenente i marginali di riga $f_i = \sum_j f_{ij}$, **c** il vettore dei marginali di colonna $f_j = \sum_i f_{ij}$, $\mathbf{D}_I$=diag(**r**) e $\mathbf{D}_J$=diag(**c**).

Per valutare l'influenza delle *J* categorie di *y* sulle *I* categorie della variabile *z*, l'ANSC trasforma la matrice **F** nella matrice $\tilde{\mathbf{F}}$ centrata rispetto all'ipotesi di indipendenza e di elemento generico:

$$\tilde{f}_{ij} = f_{ij} - f_{i.}f_{j.} \qquad (1)$$

L'attenzione viene, dunque, spostata sulla matrice centrata dei profili colonna, $\tilde{\mathbf{F}}\mathbf{D}_j^{-1}$, che considera le distribuzioni condizionate di *I* rispetto a *J*.

Geometricamente, l'ANSC ha come obiettivo la rappresentazione delle *I* categorie della variabile di risposta in un sottospazio di R^J, assumendo una metrica euclidea e un sistema di pesi pari a $\mathbf{D}_j$. Analogamente le *J* categorie della variabile esplicativa vengono rappresentate in un sottospazio di R^I, assumendo una metrica euclidea ponderata e un sistema di pesi unitario. L'ANSC, in particolare, cerca di visualizzare la dipendenza di *I* da *J* in un sottospazio di dimensioni ridotte R^m, con $m^* < m = [\min(I, J) - 1]$.

Volendo effettuare un paragone con l'Analisi delle Corrispondenze (AC), si può vedere l'ANSC come un'Analisi in Componenti Principali (ACP) sulla tripletta $\tilde{\mathbf{F}}\mathbf{D}_j^{-1}, \mathbf{I}, \mathbf{D}_j$, dove **I** è la matrice identità, mentre è noto che l'AC è un'Analisi in Componenti Principali sulla

tripletta $\tilde{\mathbf{F}}\mathbf{D}_j^{-1}, \mathbf{D}_I^{-1}, \mathbf{D}_j$. Entrambe le analisi sono, dunque, delle ACP sulla stessa matrice di partenza, con lo stesso sistema di pesi, ma in una differente metrica.

Dal punto di vista matematico, la differenza è nella decomposizione in valori singolari (SVD) che nell'ANSC viene effettuata imponendo vincoli differenti sugli autovettori di sinistra. Infatti, nell'ANSC si effettua la SVD di:

$$\tilde{\mathbf{F}}\mathbf{D}_j^{-1} = \mathbf{U}\mathbf{L}\mathbf{U}, \tag{2}$$

con i vincoli di ortonormalizzazione $\mathbf{U'U} = \mathbf{V'D}_j\mathbf{V} = \mathbf{I}$. Λ è la matrice diagonale che ha come elemento generico la radice quadrata degli autovalori, λ_α (con α =1,..,m), della matrice $\mathbf{A}=\tilde{\mathbf{F}}\mathbf{D}_j^{-1}\tilde{\mathbf{F}}'$.

Le coordinate fattoriali sull'α–esimo asse fattoriale in R^I sono così calcolate:

$$\psi_\alpha = \sqrt{\lambda_\alpha u_\alpha} \tag{3}$$

mentre in R^J sono date da:

$$\varphi_\alpha = \sqrt{\lambda_\alpha \mathrm{D}_j^{-1/2} v_\alpha} \tag{4}$$

La traccia della matrice **A**, diagonalizzata nell'analisi, rappresenta il numeratore dell'indice di predittività, caratteristico di quest'analisi, che è il τ di Goodman e Kruskal (1954):

$$\tau = \frac{\sum_\alpha \lambda_\alpha = \left(\sum_i \sum_j f_y^2 / f_i - \sum_i f_i^2 \right)}{\left(1 - \sum_i f_i^2 \right)} \tag{5}$$

Tale indice che presenta al denominatore la misura di eterogeneità del Gini per le distribuzioni condizionate (Light e Margolin, 1971), interpreta la bontà dell'analisi in termini di predittività della variabile dipendente dalla variabile esplicativa. È evidente, che in situazioni in cui il numero di righe e di colonne della matrice di partenza è elevato, caso piuttosto frequente quando si lavora con tabelle lessicali, tale indice risulta piuttosto basso.

Le rappresentazioni grafiche

Uno dei motivi principali della diffusione dell'analisi delle corrispondenze è la capacità di esprimere i risultati sotto forma di rappresentazioni grafiche di facile comprensione.

Esistono alcune regole, in parte specifiche per l'analisi non simmetrica, che è importante richiamare per poter interpretare i grafici proposti dal nostro caso studio:

a) la dispersione della nube dei punti attorno all'origine in R^J visualizza la forza del legame di dipendenza di *I* rispetto a *J*;

b) le due nubi hanno la stessa origine, grazie all'operazione di centratura ed al sistema di pesi adottato;

c) la distanza di un punto dall'origine, in R^J, mostra come la *i*-esima categoria sia influenzata dalle J categorie della variabile esplicativa, così come la distanza di un punto dall'origine, in R^I, visualizza la sua influenza sull'insieme delle *I* categorie della variabile dipendente;

d) se due punti sono vicini, in R^J, vuol dire che sono influenzati dalla stessa categoria della variabile esplicativa; se due punti sono vicini, in R^I, vuol dire che influenzano nello stesso modo le categorie della variabile di risposta;

e) la posizione relativa di una categoria della variabile di risposta e di una categoria della variabile esplicativa, può essere valutata solo in termini di coseni fra l'angolo formato dai vettori che congiungono i due punti dall'origine: un coseno grande mostra una forte influenza, viceversa un coseno piccolo mostra una bassa influenza.

3. La struttura dei dati e la strategia di analisi

Con l'obiettivo di voler studiare il linguaggio che le aziende utilizzano sui siti internet per cercare nuovo personale, sono state raccolte le informazioni disponibili nel sito www.carrierain.it. Tale portale è stato creato dall'Associazione Mercurius di Torino per agevolare i neo-laureati e i diplomati in cerca di prima occupazione nella fase di contatto con le aziende ed, inoltre, per indirizzare coloro che, già occupati, siano alla ricerca di nuovo lavoro. Il Network Mercurius è costituito complessivamente da cinque portali tematici che riguardano la formazione, il lavoro interinale, la ricerca di lavoro.

Nella sezione "Profili aziendali" del sito sono contenuti i profili di 282 aziende (incluse le società di selezione e le società di lavoro interinale). Per il caso studio proposto, si è scelto di non considerare le società di selezione e quelle di lavoro interinale, poiché hanno una funzione mediatrice tra azienda e lavoratore ed utilizzano quindi un linguaggio molto standardizzato. Sono state, dunque, considerate 167 aziende, di diversa dimensione e settore di attività, distribuite su tutto il territorio nazionale.

Scopo preciso dell'analisi è evidenziare come il linguaggio utilizzato dalle aziende per descrivere se stesse e la loro *mission*, influenzi il linguaggio da loro stesse adoperato per descrivere le diverse posizioni lavorative richieste. In una prima fase, quindi, sono stati creati due *corpora* distinti, uno ("chi siamo") contenente le informazioni generali riguardanti l'azienda (storia, struttura, attività) e un altro ("chi cerchiamo") contenente le informazioni specifiche relative alle posizioni lavorative (formazione, competenze, attitudini). I *corpora* così ottenuti sono stati poi normalizzati e lessicalizzati con procedure automatiche, al fine di rendere più facilmente confrontabili le forme costituenti i vocabolari ed evidenziare la presenza di poliformi e polirematiche di interesse per l'analisi.

In una fase successiva è stato costruito nuovamente il vocabolario dei due *corpora* con le forme normalizzate, le polirematiche e i poliformi. Si è scelto, inoltre, di introdurre un filtro sulle forme (numero di occorrenze maggiore di 2 e numero di caratteri maggiore di 4) ed effettuare quindi una lemmatizzazione interna (Lebart *et al.*, 1998), mantenendo separate solo quelle forme che si prestavano ad una interpretazione ambigua, per non perdere informazioni interessanti.

Le tabelle lessicali ottenute si presentano come due matrici di intensità: una (**Y**) in cui sono raccolte 375 forme lemmatizzate del "chi siamo" per le 167 aziende e l'altra (**Z**) in cui sono raccolte 530 forme lemmatizzate del "chi cerchiamo".

A partire da queste due matrici, si vuole costruire una matrice **F** che abbia in colonna le *J* categorie della variabile esplicativa (in **Y**), in riga le *I* della variabile dipendente (in **Z**), e il cui elemento generico consista nel numero di volte in cui ciascuna forma *i*-esima e ciascuna forma *j*-esima siano state utilizzate simultaneamente nel collettivo. La tabella lessicale **F**, quindi, che si vuole ottenere, è del tipo forme/forme e differisce evidentemente dalla classica matrice di contingenza introdotta nel paragrafo precedente. A tal fine è necessario trasformare le matrici **Y** e **Z** in matrici booleane di presenza(1)/assenza(0) di ciascuna forma nei due *corpora*. Si effettua, quindi, il loro prodotto interno **F=Z'Y**.

La matrice **F** così ottenuta è, dunque, una particolare tabella di contingenza e, per la precisione, è una *matrice di co-presenza*. Si noti che i marginali di riga e di colonna di **F** indicano il numero di volte che ciascuna forma di un *corpus* è utilizzata in combinazione con tutte le altre forme appartenenti all'altro *corpus*.

Sulla matrice **F** di *co-presenza* si propone di applicare l'ANSC, introdotta nel paragrafo precedente, con l'obiettivo di studiare il legame di dipendenza del *corpus* del "chi cerchiamo" da quello del "chi siamo". La trasformazione effettuata per passare dalle singole matrici **Z** ed **Y** alla matrice **F** ha causato però la perdita di un'informazione che nell'analisi delle tabelle lessicali risulta particolarmente importante, e cioè il numero di volte che ciascuna forma è stata usata nei due *corpora*, indicazione che non ritroviamo più sui marginali della **F**.

Per tale motivo, si introduce nell'analisi un ulteriore sistema di pesi che riguarda la variabile esplicativa e, quindi, le forme del vocabolario del *corpus* contenente le descrizioni delle aziende. Tale sistema di pesi viene definito a partire dal calcolo del *term frequency* (TF, Salton e Buckley, 1988) per ogni forma della tabella **Y**:

$$\mathrm{TF} = 0,5 + 0,5\frac{nterm_j}{\max nterm} \qquad (6)$$

dove $nterm_j$ rappresenta la frequenza all'interno del *corpus* della forma j-*esima*, mentre max *nterm* è la frequenza della forma che nel *corpus* occorre maggiormente.

In generale, in corrispondenza di livelli più alti del TF si individuano le forme con un contributo informativo maggiore in relazione alla descrizione del testo. Nell'analisi proposta si è scelto di ponderare i dati con il reciproco del TF, in modo da dare un'importanza maggiore alle forme che occorrono di meno e una minore alle forme con occorrenza più alta e, inoltre, vista l'ipotesi di relazione asimmetrica delle variabili, si è deciso di introdurre tale peso solo per la variabile esplicativa.

Nella formula (2) si introduce la matrice $\mathbf{D}_Y^{-1}$ che ha come elemento generico i pesi calcolati (6) per le forme del corpus del "chi siamo". Le formule degli autovettori e delle coordinate sono quindi, modificate tenendo conto di questo nuovo sistema di pesi.

4. I principali risultati

La matrice di co-presenza **F** analizzata ha 530 righe (le parole del *corpus* del "chi cerchiamo") e 375 colonne (le parole del *corpus* del "chi siamo ").

L'ANSC è sta effettuata considerando il sistema aggiuntivo dei pesi ($\mathbf{D}_Y^{-1}$). L'indice di predittività ottenuto è pari a 0,0041. Se non avessimo utilizzato alcun sistema di pesi aggiuntivo, il τ sarebbe stato pari a 0,0022, mentre se avessimo considerato come pesi la frequenza di ogni singola forma sul totale di forme utilizzate ($nterm_j / \sum_j nterm_j$) il τ sarebbe stato pari a 0,0016. Il valore dell'indice di predittività della nostra analisi è un primo segnale che il sistema di pesi introdotto migliora i risultati dell'analisi.

Nella tavola 1 è riportata la percentuale spiegata della struttura di dipendenza tra il "chi siamo" e il "chi cerchiamo", dai primi 10 autovalori. La percentuale del τ di Goodman-Kruskal spiegata dal primo piano fattoriale è pari al 13,4%; tale valore, sebbene apparentemente basso, è in realtà significativo considerando che ogni forma singolarmente spiega solo lo 0,22%.

autovalore	% spiegata	% cumulata
λ_1	8,3	8,3
λ_2	5,1	13,4
λ_3	4,5	17,9
λ_4	4,3	22,1
λ_5	3,7	25,8
λ_6	3,0	28,8
λ_7	2,8	31,6
λ_8	2,7	34,3
λ_9	2,7	37,0
λ_{10}	2,4	39,4

Tavola 1. Percentuale della struttura di dipendenza spiegata dai primi 10 autovalori

Per interpretare i risultati delle rappresentazioni grafiche ottenute, è opportuno ricordare che, nell'ANSC, occorre mantenere separate le rappresentazioni nei due spazi e valutarle congiuntamente alla luce della regola *e* del paragrafo 2.

Sul piano fattoriale dove sono rappresentate le forme del "chi siamo" (Fig. 1a), troviamo alla sinistra del primo asse quelle forme che, caratterizzate da coordinate negative e contributo assoluto alto (vedi Tav. 2), individuano aziende con attività tradizionali (*costruzione, raffinazione, elettrica, petrolifera, commercializzazione, etc.*). Diversamente, a destra si evidenziano forme che caratterizzano aziende operanti nei settori innovativi dell'informatica e dei servizi web (*hardware, software, networking, etc.*).

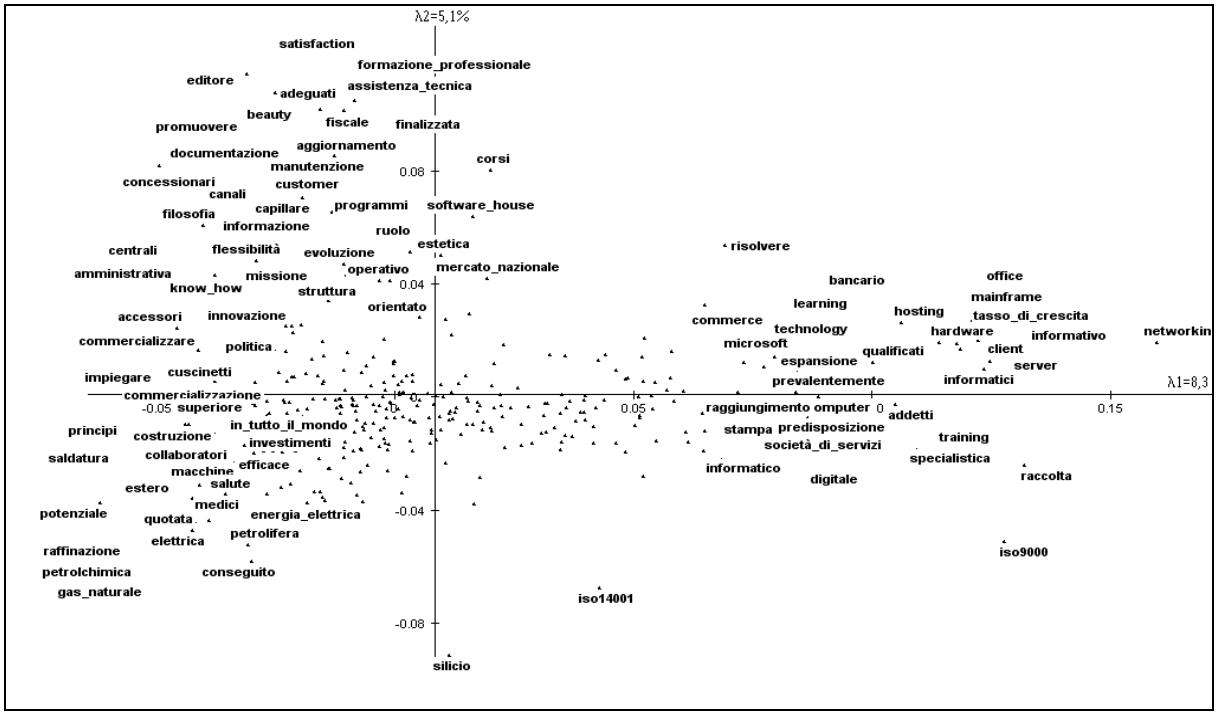


Figura 1a. ANSC primo piano fattoriale della rappresentazione delle parole del "chi siamo"

	forme	coordinata	contributo assoluto		**forme**	coordinata	contributo assoluto
1	potenziale	-0,062059434	4,04E-06	1	networking	0,159636049	1,08E-05
2	gas_naturale	-0,060309706	1,85E-06	2	raccolta	0,131902066	2,28E-06
3	raffinazione	-0,058871858	3,08E-06	3	iso9000	0,127858346	3,15E-06
4	petrolchimica	-0,058304691	3,02E-06	4	client	0,124662377	1,07E-05
5	impiegare	-0,053683353	6,63E-06	5	server	0,123559133	9,43E-06
6	documentazione	-0,049543869	3,58E-06	6	office	0,123492745	6,59E-06
7	accessori	-0,045803329	1,64E-06	7	tasso_di_crescita	0,122193708	3,11E-06
8	principi	-0,044076608	4,03E-06	8	mainframe	0,121007841	5,31E-06
9	collaboratori	-0,04398157	2,11E-06	9	informatici	0,118569509	1,22E-05
10	costruzione	-0,043271691	3,09E-06	10	informativo	0,117866419	8,90E-06
11	quotata	-0,042630399	2,36E-06	11	training	0,115401966	3,39E-06
12	elettrica	-0,042575964	1,92E-06	12	hardware	0,114178771	1,46E-05
13	petrolifera	-0,042111233	2,90E-06	13	specialistica	0,109334775	3,51E-06
14	politica	-0,041782517	3,23E-06	14	hosting	0,106078927	5,04E-06
15	commercializzare	-0,041414147	2,66E-06	15	addetti	0,104814214	2,29E-06
16	Estero	-0,040995457	2,34E-06	16	qualificati	0,100246884	5,51E-06
17	superiore	-0,040982496	2,17E-06	17	computer	0,088871937	7,68E-06
18	Filosofia	-0,040400157	3,30E-06	18	bancario	0,087550783	3,67E-06
19	Energia_elettrica	-0,03904156	1,92E-06	19	società_di_servizi	0,086850543	3,55E-06
20	saldatura	-0,039004015	2,70E-06	20	stampa	0,086635692	1,97E-06
21	in_tutto_il_mondo	-0,03816545	3,98E-06	21	learning	0,084430836	3,14E-06

Tavola 2. Coordinate e contributi sul primo asse fattoriale

Il secondo asse fattoriale (vedi Tav. 3 per le coordinate) contrappone dal basso verso l'alto le forme riferite ad aziende manifatturiere di grandi dimensioni con mercato internazionale (*quotata*, *estero*, *in tutto il mondo*, etc.) a quelle relative ad aziende di servizi (*fiscale*, *manutenzione*, *amministrative*, *concessionari*) che si collocano sul mercato nazionale (*nazionale*) e hanno una strategia di mercato orientata al cliente (*customer satisfaction)*.

	forme	coordinata	contributo assoluto		**forme**	coordinata	contributo assoluto
1	silicio	-0,091535061	1,55E-06	1	networking	0,159636049	1,08E-05
2	iso14001	-0,067615654	1,38E-06	2	raccolta	0,131902066	2,28E-06
3	gas_naturale	-0,059334586	1,79E-06	3	iso9000	0,127858346	3,15E-06
4	raffinazione	-0,055367507	2,72E-06	4	client	0,124662377	1,07E-05
5	petrolchimica	-0,054834102	2,67E-06	5	server	0,123559133	9,43E-06
6	iso9000	-0,051072138	5,03E-07	6	office	0,123492745	6,59E-06

	forme	coordinata	contributo assoluto		**forme**	coordinata	contributo assoluto
7	elettrica	-0,047323764	2,37E-06	7	tasso_di_crescita	0,122193708	3,11E-06
8	petrolifera	-0,04390463	3,15E-06	8	mainframe	0,121007841	5,31E-06
9	energia_elettrica	-0,043485566	2,38E-06	9	informatici	0,118569509	1,22E-05
10	dispositivi	-0,037331704	9,68E-07	10	informativo	0,117866419	8,90E-06
11	semiconduttori	-0,036716715	1,13E-06	11	training	0,115401966	3,39E-06
12	quotata	-0,035864847	1,67E-06	12	hardware	0,114178771	1,46E-05
13	elettronica	-0,033812638	1,93E-06	13	specialistica	0,109334775	3,51E-06
14	produttori	-0,033582915	1,64E-06	14	hosting	0,106078927	5,04E-06
15	ingegneria	-0,033459492	1,89E-06	15	addetti	0,104814214	2,29E-06
16	estero	-0,030920056	1,33E-06	16	qualificati	0,100246884	5,51E-06
17	acciaio	-0,026195629	6,14E-07	17	computer	0,088871937	7,68E-06
18	industria	-0,025275743	1,95E-06	18	bancario	0,087550783	3,67E-06
19	automobilistiche	-0,025060031	3,49E-07	19	società_di_servizi	0,086850543	3,55E-06

Tavola 3. Coordinate e contributi sul secondo asse fattoriale

Passando alla lettura della rappresentazione della variabile dipendente "chi cerchiamo" (Fig. 1b), è possibile leggere la dipendenza di tali forme da quelle precedenti, in termini di coseni di angoli. Come si evince dai risultati la relazione è chiaramente confermata.

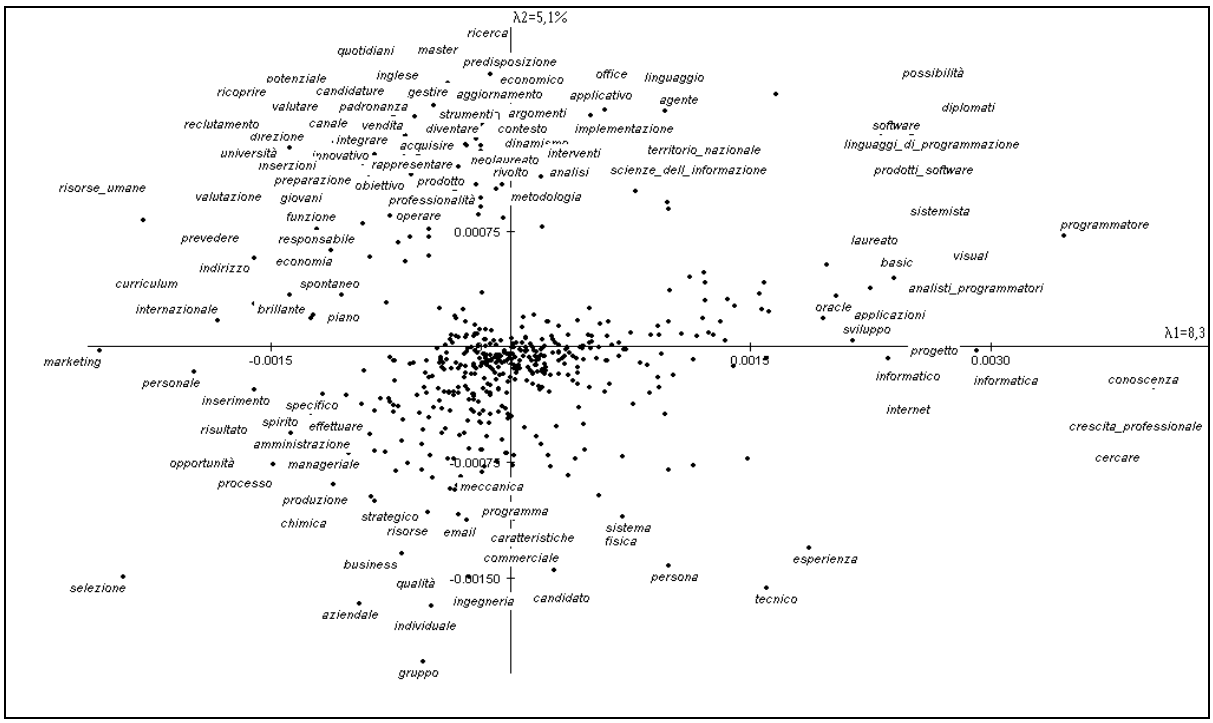


Figura 1b. ANSC primo piano fattoriale della rappresentazione delle parole del "chi cerchiamo"

In alto a sinistra dove, nel grafico precedente erano rappresentate le forme raffiguranti le imprese di servizi tradizionali, orientate al cliente, qui troviamo forme quali *master*, *inglese*,

aggiornamento, *preparazione*, che descrivono il profilo di un candidato che si adatta bene alla tipologia di azienda considerata. In basso a sinistra dove, nel grafico precedente trovavamo forme raffiguranti le multinazionali manifatturiere, qui troviamo forme quali *marketing*, *personale*, *amministrazione*, *produzione*, *commerciale*, che descrivono chiaramente la funzione aziendale che i candidati andranno a svolgere in quelle aziende.

In alto a destra dove, come visto, erano rappresentate le forme raffiguranti le imprese informatiche e le imprese di servizi innovativi, vi sono forme quali *programmatore*, *linguaggi di programmazione*, *sistemista*, *diplomato*, *laureato*, tipiche dello specifico profilo richiesto.

5. Conclusioni

In questo lavoro si è voluto proporre una strategia di analisi delle relazioni di dipendenza tra *corpora* testuali a partire dalla definizione di una particolare tabella lessicale del tipo forme/forme. L'utilizzo di tale matrice che trae origine da due *corpora* rilevati sulle stesse unità ma relativi a documenti distinti, consente evidentemente di generalizzare la strategia a casi studio piuttosto frequenti, come ad esempio quelli in cui i documenti sono relativi ad occasioni differenti. Si pensi, ad esempio, alle indagini ripetute nel tempo, dove evidentemente l'obiettivo può essere quello di valutare la relazione tra le risposte date dagli stessi soggetti in tempi differenti, sullo stesso argomento.

Un interessante sviluppo dell'analisi è quello di prevedere la possibilità di inserire anche le eventuali informazioni disponibili sulle unità che, invece, nell'approccio proposto non sono state considerate per considerare, invece la dipendenza del *corpus* del "chi cerchiamo" da quello del "chi siamo".

Si può, quindi, pensare di recuperare le informazioni sulle unità proiettandole in supplementare attraverso un operatore di proiezione che tenga conto del fatto che il sottospazio di riferimento è definito dalle distribuzioni condizionate delle forme testuali. Inoltre, si può pensare di sfruttare le potenzialità dell'analisi simbolica dei dati per rappresentare sia categorie di individui come oggetti simbolici definiti dalle modalità delle forme per ciascun *corpus*, sia le intensità della matrice di *co-presenza* come oggetti simbolici definiti dalle caratteristiche di ciascuna cella. Quest'ultimo obiettivo, evidentemente, richiede la definizione di particolari distribuzioni di probabilità delle variabili considerate, spostando l'attenzione sull'eventualità di introdurre una possibile modellizzazione dei dati.

Bibliografia

Balbi S. (1995). Non symmetrical correspondence analysis of textual data and confidence regions for graphical forms. In *JADT 1995*, vol (2) : 5-12

Goodman L.A. e Kruskal W.H. (1954). Measures of association for cross-classification. *J.A.S.A*, vol (49) : 732-764.

Lauro N. e D'Ambra L. (1984). L'analyse non symétrique des correspondances. In Diday *et al.* (Eds), *Data Analysis and Informatics*. NH : 433-446.

Lebart L., Salem A. e Berry L. (1991). Recent developments in the statistical processing of textual data. *Applied Stochastic Models and Data Analysis*, vol (7) : 47-62.

Lebart L., Salem A. e Berry L. (1998). *Exploring textual data*. Kluwer Academic Publisher.

Light R.J. e Margolin B.H. (1971). An analysis for variance for categorical data. *J.A.S.A.*, vol. (335/66) : 534-544.

Salton G. e Buckley C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, vol (24/5) : 513-523.

Text Categorisation of Racist Texts Using a Support Vector Machine

Edel P. Greevy[1], Alan F. Smeaton[2]

[1]PRINCIP Project – SALIS – Dublin City University – Dublin 9 – Ireland
[2]Centre for Digital Video Processing – Dublin City University – Dublin 9 – Ireland

Abstract

The automatic processing of text is a major challenge because of the increasing availability of textual information and the need to organise and manage such information effectively and efficiently. Automatic Text Categorisation is one of a number of functions we would like to have available to us and involves the assignment of one or more predefined categories to text documents in order that they can be effectively managed. In this paper we examine the problems associated with categorising texts documents (web pages) based on whether or not they are racist. We describe work in the PRINCIP project, which aims at the development of a system to detect racism based on the results of linguistic and statistical analysis of candidate texts. We take what we have learned from the PRINCIP research and apply machine learning techniques, specifically Support Vector Machines, to automatically categorise web pages. Our work shows that it is possible to develop automatic categorisation of web pages, based on these approaches.

Keywords: machine learning, text categorisation, support vector machines.

1. Introduction

Automatic Text Categorisation (TC) is the task of assigning predefined categories to free text documents (Yang, 1999). Texts are assigned to categories based on a confidence score that is suggested by a training set of labelled documents. This confidence score usually ranges between 0 and 1 and in order to arrive at a yes/no decision for the inclusion/exclusion of a document in a category, the confidence score is mapped onto one of the Boolean values {0,1} using thresholds. TC techniques have been successfully applied to many domains: for the classification of news stories into relevant newsgroups such as sports, politics, environmental issues, to detect spam and to assign web pages into Yahoo!-like web directories.

The PRINCIP project aims at the realisation of a multilingual system for detecting racist documents on the web. PRINCIP is primarily a linguistics-based project working to establish common methodologies across three languages (French, German and English) through the corpus-based analysis of racist content with the aim of building a linguistic knowledge base (KB). Tagging, parsing, linguistic and statistical analysis using bespoke and existing linguistic tools and software suites, all play major roles in building this knowledge base. The linguistic patterns identified during analysis of web pages can be formulated into rules and used in a categorisation system, to allow for detection of illicit content on the web.

In this paper we present an overview of the techniques we use to automatically develop and evaluate a text categorisation system in PRINCIP. Our approach is dictated by the results of linguistic and statistical analysis of the text appearing on web pages, experiments conducted during the PRINCIP project. For the purpose of this study, a Support Vector Machine (SVM), which is a machine learning method, will be trained on several representations of the datasets collected for the PRINCIP project.

2. Racism on the Web and Ways to Detect It

In this paper we are concerned with racism in English-speaking countries only. The UK and Ireland are the main English-speaking countries in Europe while the USA, Canada, Australia, New Zealand, South Africa are the main countries outside of Europe. Those countries in which English is a strong second language include Israel, Holland and the Scandinavian countries as well as many other places in Europe. Thus we can see that racism in English can originate in many corners of the globe. Targets of racism differ from country to country, with, for example, immigrants, refugees and non-nationals being strong targets in Ireland; African-Americans, Hispanics and Asians being likely targets in the USA while Aborigines are the main target of racism in Australia. It must also be noted that the targets of racism constantly shift and change. September 11th, the ongoing war in Israel and Palestine, the war in Iraq, American foreign policy, the killings of white farmers in South Africa – current affairs – shift the attention from one group to another. Our corpus of web documents was gathered in September 2002 and it is already somewhat out of date as a result of the world events since then.

Internet legislation also impacts the presence of racism on the web. This differs across the globe with the U.S. being one of the most liberal. The U.S. First Amendment entitles its citizens to the right to freedom of speech and for that reason the majority of racism found on the WWW is US-based, though some groups, for strategic reasons, take advantage of more restrictive laws (e.g. in Canada) as it warrants them more publicity.

Non-technological methods that exist for the detection and removal of hate online include the setting up of regulatory authorities. For example the Netherlands has set up the Complaints Bureau for Discrimination; hotlines exist in EU countries which allow for potential breaches of legislation to be reported. Sites are investigated and if found to be illegal, are eventually removed. Such solutions are found to be weak because of the fluidity and size of the Internet. Documents originating in the USA, where legislation is most liberal, can be accessed across the globe but belong to another jurisdiction. Technical approaches thus far implemented include Internet Content Filters or Label Bureaus, which simply label sites and filter offensive ones (Internet Content Rating Association). Email is typically filtered using regular expressions containing keywords but this approach is unreliable, as it will only filter those emails containing known keywords. The Safer Internet Action Plan, which is sponsored by the European Commission, is currently funding various filtering and rating projects some of which include: the ICRAsafe project which will create a system to allow responsible adults to restrict children's access to Internet content that may harm them or which is otherwise considered undesirable by the adult; NETPROTECTII is a European tool for Internet access filtering to provide textual filtering in eight European languages. All project descriptions can be viewed under the URL provided for the Safer Internet Action Plan.

Current methods of filtering racism rely heavily on either keywords or the labelling of offensive material. In order to implement successful systems, a considerable human effort is required, not only in the initial stages of filter construction but also in an ongoing basis as the targets of racism change, the language evolves, existing websites are edited or new websites are added. Automatic text categorisation techniques are reported to have been successful when applied to other domains such as news story categorisation or the categorisation of web pages into Yahoo!-like directories, with results comparable to human evaluation and performance. Such methods lead to vast improvements in productivity as well as savings in terms of time and manpower, as the same human effort is not required. Given racism on the web changes so rapidly, it is one area that may benefit from the application of automatic techniques to text categorisation.

3. Text Categorisation

Text Categorisation (TC) is concerned with the automatic assignment of documents to predefined categories. TC is traditionally a content-based management task and has much in common with it's neighbour Information Retrieval (IR), borrowing and applying much of the basic IR techniques. IR is concerned with the matching of a user's information need, expressed as a query, against a corpus of documents in order to rank documents in the corpus in order of their estimated relevance to the information need. As the field of TC has progressed it is no longer just concerned with the assignment of documents into categories such as sport, politics or environmental issues but has attempted to solve more complex tasks such as the classification of documents according to genre (Finn and Kushmerick, 2003; Finn *et al.*, 2002). In recent years machine learning techniques have been applied to text categorisation problems. The promise of a future of machines capable of reading, examining and making decisions about free text has generated more interest in the field.

There are three main processes involved in text categorisation, namely:

1. Indexing and other pre-processing techniques are performed on initial training and test corpora and also on documents to be categorised by the classifier when in operation.
2. Classifiers take the training data as input and learn features of that data so that when presented with unseen data, it will use those learned features to make a decision about the category in which the document is assigned.
3. An evaluation procedure is used to measure the effectiveness of the classifier. After being trained on the training data, the classifier is presented with test data and the performance of the classifier is evaluated using precision and recall.

3.1. *Indexing and Pre-processing Operations*

Because text documents are not interpretable by a classifier, an indexing procedure must be applied to the text so as to map it onto an appropriate representation that can be fed to the classification system. Almost all representations consist of a set of *terms* each of which may be assigned a *weight* in each document to reflect its degree of importance for that document. Differences between various indexing approaches are reflected by different interpretations of what constitutes a term and of how weights are measured in a document.

3.1.1. *What can a term be?*

Term generation varies in the amount of linguistic and statistical sophistication used. To form the simplest indexing language each word can be treated equally as a feature. This is a common approach referred to as the *bag-of-words (BOW)* approach. However, relationships such as polysemy and synonymy which exist between words, can lead to many errors. For this reason, more complex methods are investigated for the creation of an effective set of *terms.* Where phrases are used as indexing terms, phrases can be determined using linguistic information (i.e. identified according to the grammar of the language) or by using statistical methods (i.e. identified according to the recurring frequency of a set of words).

The application of linguistic procedures to a text allows us to express language in terms of the roles words play in a text and the relationships between words. This, in turn, provides a classifier with richer information about a document. Experiments conducted using linguistic information are inconsistent with some findings revealing they do not perform as well as BOW (Lewis, 1992; Smeaton, 1997) while others disagree (Fürnkranz *et al.*, 1998).

3.1.2. *Using Linguistic Information in Term Generation*

Some of the linguistic approaches to generating index terms reported in the literature include:

- Chandrasekar and Srinivas illustrated how syntactic information can be effective in filtering out irrelevant documents after documents have been retrieved by a search engine. They reported on the performance of two different methods of syntactic labelling namely Part of Speech (POS) Tagging and Supertagging (Chandrasekar and Srinivas, 1998).
- Fürnkranz *et al.* illustrated how the use of linguistic phrases as input features can improve precision at the expense of recall. Linguistic phrases were constructed using a system called AUTOSLOG which is an automatic method for extracting patterns from a POS tagged text. The system is fed noun phrases which are used in the construction of linguistic patterns which are in turn used by the classifier during the categorisation of documents (Fürnkranz *et al.*, 1998).
- Lewis performed tests on the use of syntactic indexing phrases, clustering of these phrases and clustering of words and found these approaches to be less effective than the frequency of occurrence of words. Lewis proved word-based features to be more effective and attractive since a greater effort, both computationally and time-wise, is required in order to build a feature set of phrases (Lewis, 1992).
- In their study on the application of TC techniques to classify documents along dimensions that are orthogonal to topic, for example whether a document is primarily facts or an expression of someone's opinion or whether a document is positive or negative, Finn, Kushermick and Smyth found the distribution of POSs to be more effective than the BOW approach (Finn and Kushmerick, 2003; Finn *et al.*, 2002).

3.1.3. *Using Statistics in Term Generation*

The following illustrate the kinds of statistics-orientated approaches to the generation of indexing terms that have been explored in the literature.

- Mladenic and Grobelnik combine feature generation with feature selection through the use of statistical methods. What is interesting about this method is that the BOW vector space is enriched with phrasal information (word sequences of between 2 and 5 characters), meaning the classifier is not relying on the performance of phrases alone. Word sequences of size 3 proved to be the most effective (Mladenic and Grobelnik, 1998).
- Typically classifiers assume that the context of a word *w* has no impact on the meaning of *w,* which is of course not true. For this reason Cohen and Singer aim to construct a classifier that allows the context of a word *w* to affect whether the presence or absence of *w* will contribute to a classification. Cohen and Singer investigated two algorithms each of which have different notions as to what constitutes context. For the RIPPER algorithm a context of a word *w* is interpreted as a number of other words that must co-occur with *w,* where order and location in the document are irrelevant. Sleeping-experts, on the other hand, interprets the context of a word *w* as consisting of words that occur near *w* and in a fixed order (Cohen and Singer, 1998).
- Fürnkranz employed an algorithm very similar to that used by (Mladenic and Groblenik, 1998) for the generation of n-grams. Each document represented *m* set-valued features, one for each n-gram size $1<m<MaxNGramSize$ meaning for example that when $m=3$ all 3-grams, 2-grams and 1-grams are included in the feature set (Fürnkranz, 1998).
- Tan *et al.* investigated the use of bigrams to enhance text classification. In this experiment Tan *et al.* used bigrams in addition to unigrams. Those unigrams that appeared in a significant number of documents were selected and used as seeds for the generation of bigrams. Bigrams were generated and chosen on the basis that at least one of the pairs of bigrams had to be a seed (Tan *et al.*, 2002).

3.1.4. *Other Pre-processing Operations*

Before a document is indexed, the normal procedure in IR and TC is to remove *stop words*. Stop words comprise those words that are neutral to the topic of the document and would therefore generally contribute very little to the classification of a document. They are often defined by a *stoplist* and include articles, prepositions, conjunctions and some high-frequency words. This technique is performed so as to reduce the number of index terms in a document, to enhance computational efficiency and to minimise the amount of superfluous information in the term space. Different methods have been explored for the generation of stoplists. In their 1996 JASIS paper, Yang and Wilbur (Yang and Wilbur, 1996) apply the Wilbur-Sirotkin stop word identification method to text classification in order to reduce the computational cost without having to trade off on categorisation effectiveness (Wilbur and Sirotkin, 1992).

Stop words are not removed in experiments using syntactic information as terms (Lewis, 1992; Chandrasekar and Srinivas, 1998). Such experiments require the presence of all words in a sentence or document in order to assign the correct POS tags. *Term* or *feature extraction* techniques are used instead to identify the most discriminating and effective patterns.

3.2. *Post-Indexing Operations*

3.2.1. *Assigning weights to terms*

A term can be weighted using binary weights i.e. 1 denotes presence and 0 denotes absence, or using frequency weights. More complex term weighting methods exist, with weights usually ranging between 0 and 1. The *TF*IDF* term weighting function (Salton and Buckley, 1988) is a commonly used method.

3.2.2. *Dimensionality Reduction (DR)*

In TC since the number of terms occurring just once in a corpus can be extremely high, in some cases, efforts are made to reduce the dimensionality of the term space from r to $r`$. Large vector spaces can be problematic and can lead to *overfitting.* A good example of overfitting as provided by (Sebastiani, 2002) is that of a classifier trained on three examples for the category CARS FOR SALE. Two of the advertisements were concerned with the sale of blue cars and therefore the classifier considered the colour of the car (i.e. blue) to be a characteristic of the category. In other words classifiers affected by overfitting tend to be exceedingly good at classifying the training data but not so good at classifying unseen data.

There are two main approaches to DR, namely *term selection* and *term extraction.* In term selection, the $r`$ terms are chosen by selecting a subset of the original $r$ terms without loss in effectiveness. Document Frequency, Information Gain, Mutual Information, Chi-square and Term Strength are popular methods for term selection. In term extraction, the $r`$ terms that are extracted may not at all resemble the original r terms. Rather the $r`$ terms are obtained through a series of alterations, combinations, transformations etc. of the original r terms. Term Clustering and Latent Semantic Indexing are popular methods for term extraction.

4. Machine Learning

There is no conventional algorithm for the task of assigning a document to a predefined category, as no accurate mathematical model of the solution exists. Given a set of examples, we might be able to define input and output values for each example but we are unable to define how, given a certain input we arrive at the desired output. The relationship between the input and desired output is too complex to be captured in an algorithm and so the only way such a problem can be dealt with is by using machine-learning techniques.

Machine learning (ML) can be broadly split into two main areas: supervised learning and unsupervised learning. In supervised learning, the machine knows the output of an input pattern and tries to learn patterns that would arrive at the desired output. In unsupervised learning, the training set consists of input patterns only and the machine is trained without having any prior knowledge of the output. Its task is to learn to adapt based on the experiences of the previous training patterns.

Binary classifiers have a binary output i.e. *{0, 1}* meaning a document either belongs to a category or it does not. Multi-class classifiers allow a document to be categorised in one of a finite number of categories. In regression models the output is a real numbered output. ML has been applied to a wide range of areas from speech recognition, hand-written character recognition, image detection, POS tagging to medical diagnosis and prognosis and even learning to fly.

4.1. *Some Approaches to Classifier Construction*

Many methods, approaches and algorithms exist for the construction of a text classification system. Some of the more popular approaches are mentioned below.

Probabilistic classifiers view the classification problem in terms of a probability that a document *D* of binary or weighted terms belongs to a category *C*. The probability is calculated by applying Bayes Theorem. Many practitioners have experimented with probabilistic classifiers in the literature (Lewis and Ringuette, 1994; Yang and Liu, 1999; Chai *et al.*, 2002; Mladenic and Grobelnik, 1998; Joachims, 1998).

A decision tree classifier consists of a tree where each internal node is labelled by a term used to test an attribute. Each branch corresponds to an attribute value representing the weight that a term has in a test document and each leaf node is labelled by a category which is used to assign a classification. A document *d* is categorised by recursively testing for the weights that the terms have in the representation of *d*. This step is repeated until a leaf node is reached where the label of the leaf node i.e. the category is then assigned to *d.* (Lewis and Ringuette, 1994; Goller *et al.*, 2000; Joachims, 1998)

Decision rules first of all create a dictionary containing the features or attributes that represent individual documents in a collection or domain. A representation maps each individual document in a training set using the dictionary. Each document is assigned a label that denotes which category it belongs to. The objective is to find sets of decision rules or patterns that distinguish one category from the others (Apté *et al.*, 1994).

The Rocchio method is a vector-space-method which is very often used in information retrieval for relevance feedback, document filtering and routing. A prototype vector (centroid) is created for each category using a training corpus and is in effect the average of all positive examples. Document vectors belonging to a category are weighted positively and other documents are weighted negatively. The Rocchio classifier rewards the closeness of a test document to the centroid of the positive training examples and its distance from the centroid of the negative training examples (Goller *et al.*, 2000; Drucker *et al.*, 2001; Joachims, 1998).

A neural network classifier consists of a network of units. Input units represent terms and output units represent categories and they are connected by edges that have weights, which represent the conditional dependence relations between the I/O units. A document *d* is classified by taking its term weights and assigning them to the input units; the units are then propagated through the network and the value that the output unit takes up determines the categorisation decision (Yang and Liu, 1999).

4.2. *Support Vector Machines*

Support Vector Machines (SVMs) are one of the newer ML approaches, introduced by Vapnik in 1992. SVMs are based on statistical methods to minimise the risk of error and offer solutions to optimise generalisation performance.

One justification for using a SVM for TC is that it is "a principled and very powerful method that in the few years since its introduction has already outperformed most other systems in a wide variety of applications" (Cristianini and Shawe-Taylor, 2000). SVMs overcome the many problems associated with efficiency of training thereby making them a very attractive learning method.

SVMs are capable of overcoming the problems associated with high dimensional spaces (e.g. overfitting) due to the sophisticated statistical learning theory used. This means solutions can always be found efficiently even for training sets with thousands of examples.

The compact representation of the hypothesis being learned (in our case the categorisation of documents) means that evaluation on unseen input is very fast thereby making it efficient when it comes to testing. Within SVM terminology, *generalisation performance* refers to how well a hypothesis correctly classifies data not in the training set and a good learning machine will optimise generalisation performance. The *capacity* of a machine is the ability of the machine to learn any training set without error. A machine can have too much or too little capacity and this affects generalisation performance – too much capacity causes *overfitting.* The *VC-dimension* (Vapnik Chervonenkis) is a direct measure of the capacity of a machine.

4.2.1. *Generalisation Theory and how SVMs work*

In SVMs, the task is to learn the relationship between input/output pairs – this is known as the *target function.* When presented with an unseen document the machine can make a decision about the *target* class of the document. The *decision function* estimates the target function and this function is chosen from a set of candidate functions referred to as *hypotheses.*

N input/output pairings are represented by a vector $x_i \varepsilon R^n$, $i = 1, \ldots, n$ and the associated truth or class y_i where y_i is *1* if a document *d* belongs to category *c* and *–1* otherwise. The task of the machine is to choose the mapping $x_i \rightarrow y_i$ that minimises the risk of error.

$R(\alpha)$ is referred to as the actual risk. This cannot be computed as it depends on the unknown probability distribution $P(x,y)$ from which the data are drawn.

$R_{emp}(\alpha)$ is the empirical risk and is measured by the mean rate of error on the training set for a fixed and finite number of observations or training sets $\{x_i,y_i\}$.

The following inequality holds with probability *1-n*, if *l* is the number of training points:

$$R(\alpha) <= Remp(\alpha) + sqrt[(h(log(2l/h) + 1) – log(n/4)) / l]$$

The right hand side of this inequality, referred to as the risk bound or the VC-confidence, can be calculated if *h* is known. Minimising the risk bound or VC-confidence puts a bound on the actual risk. In this inequality equation, *h* is known as the VC-dimension and is the maximum number of training points that can be arbitrarily labelled by a set of functions. The VC-confidence increases as *h* increases so a lower VC-dimension will give a lower bound on the risk.

4.2.2. *Structural Risk Minimisation Principle*

SVMs implement the structural risk minimisation principle which attempts to overcome the problem of choosing the set of functions (i.e. the target function from all hypotheses) that has an appropriate VC-dimension while at the same time minimising the bound on the actual risk.

The entire set of functions is divided into nested subsets with decreasing capacity – for each subset either h or a bound on h is computed. This can be done by training a series of machines i.e. one for each subset. The goal of the training is to minimise the empirical risk $R_{emp}(\alpha)$ for a given subset. The subset of functions whose sum of empirical risk and VC confidence is minimal is chosen as the trained machine.

4.3. *Evaluation of Text Categorisation Systems*

Classifiers are experimentally evaluated by presenting new unseen data to the system. The efficiency of a classifier is tested by evaluating its capability of making the right categorisation decision. Precision and Recall are commonly used evaluative methods in IR and TC has borrowed these and applied them to the case of document categorisation. Precision is a measure of how accurate the system is at classifying unseen data for a particular category and is defined as the conditional probability $P(ca_{ix} = 1 \mid a_{ix} = 1)$, i.e. the probability that if a random document d_x is classified under c_i, this decision is taken. Recall is a measure of the degree of completeness or coverage for a specific category and is defined as the conditional probability $P(a_{ix} = 1 \mid ca_{ix} = 1)$, i.e. the probability that, if a random document d_x ought to be classified under c_i, this decision is taken (Sebastiani, 2002) .

5. TC in the PRINCIP Project

The PRINCIP project aims to build a system to detect and filter racism on the Internet using those rules found during linguistic analysis. In our research we use text categorisation methods to automatically achieve the same. Detecting racism on the Internet is not just a topic-based problem, rather it is more similar to genre detection as described by Finn, Kushmerick and Smyth (Finn and Kushmerick, 2003; Finn *et al.*, 2002), in that we are not really concerned with the topic itself but we are trying to identify features that will discern the author's attitude in relation to the topic, something which is orthogonal to the topic. In their work on genre detection Finn *et al.* found the distribution of POS to outperform the BOW approach for genre detection. Our own experiments in PRINCIP revealed there to be differences in some lexical, collocation and POS distributions across racist and non-racist documents (Lechleiter and Greevy, 2003; Martin, 2003a and 2003b; Gibbon and Greevy, 2003). In this study we perform a comparative analysis of the TC of racist texts by training Support Vector Machines on three representations: bag-of-words, n-gram word sequences, and POS, approaches which are primarily driven by the results of linguistic analysis in the PRINCIP project.

All three representations have already been tried and tested, both in IR and in text categorisation (Mladenic and Grobelnik, 1998; Tan *et al.*, 2002; Finn and Kushmerick, 2003; Finn *et al.*, 2002; Fürnkranz *et al.*, 1998; Lewis, 1992; Smeaton, 1997; Chandrasekar and Srinivas, 1998). However they have been examined in the context of other domains and different types of classification problems, that is, classification problems that are related to topic or content rather than to attitude or opinion. No such experiments have been conducted on the detection of racism on the Internet.

5.1. *About the dataset*

To conduct PRINCIP experiments, a web corpus of 3 million words, (approximately 500 documents per dataset) was collected. This consists of three datasets – web pages which are racist, anti-racist and neutral, i.e. neutral documents comprise those found using the same techniques used to detect racism but which are unrelated to the topic of racism. The datasets are in two formats – plain text and POS tagged.

For this study we are concerned with a binary classification problem, that is, either a document is racist or it is not. Therefore the training and test sets will contain positive examples i.e. racist texts and negative examples i.e. non-racist texts, comprising anti-racist and neutral pages. Each of the positive and negative datasets contains 500 documents.

When building the dataset, a combination of approaches was used to avoid circularity, to target a diverse collection of documents from different domains, and to target different groups. Yahoo! and Google directories were browsed. A list of potentially racist keywords and phrases was constructed. Recent and current affairs provided useful clues for the building of the list, as did studying research on racist discourse (van Dijk, 1987; Wodak and Reisigl, 2001). The list was submitted to search engines such as Google and AlltheWeb. We assumed that racist sites (and anti-racist) link to sites of a similar nature. It followed that downloading hyperlinks in a document proved a particularly useful method in corpus building.

5.2. *SVMs for PRINCIP*

Support Vector Machines were used to learn the features of the training sets and classify new unseen documents. SVMs are a very powerful learning method that "in the few years since its introduction has already outperformed most other systems in a wide variety of applications" (Cristianini and Shawe-Taylor, 2000). SVMs overcome many of the problems associated with efficiency of training such as overfitting and they are capable of generalising well in high dimensional spaces thereby making them a very attractive learning method.

5.3. *Results*

We built three representations of each dataset i.e. BOW, n-gram word sequences and POS tagged documents. The positive and negative datasets were divided into training and test sets (see table 1). The SVM learns the trainings set and uses those learned features to classify unseen documents i.e. the test set. In this study we split the training and test sets into different sizes to evaluate the impact of larger training sets on the SVM. Table 1 outlines the different size training and test sets used. Each of the experiments conducted on the different representations is evaluated in terms of precision and recall.

	Set 1	*Set 2*	*Set 3*	*Set 4*
No. Docs in Training Set	200	400	600	800
No. Docs in Test Set	60	100	150	200

Table 1. Illustrates number of documents in each set.

5.3.1. *Bag-of-Words*

The first representation used the simplest indexing language i.e. the BOW approach. Our investigation of lexical items equally consistent in each dataset revealed some words to be thirty percent more prevalent in racist texts e.g. *must, never, once, ever, same, very, course, fact, white, race, nation*. Modals, adverbs and truth claims were among this list. The use of modals, representing the taking of absolute positions, and the use of argumentation structures such as truth claims like *fact* or *of course* are both indicative of the discourse of racist language (Gibbon and Greevy, 2003; Lechleiter and Greevy, 2003). Though these same lexical items may be used by potentially anyone, the SVM results on the BOW representation prove promising for classification problem at hand.

Dataset	Term Weight	Accuracy on Test set	Precision	Recall
Set 1	Number of occurrences	60.00%	87.50%	23.33%
Set 1	Frequency	86.67%	92.31%	80.00%

Table 2. Evaluation of different methods of measuring term weight

Using set 1, we compared two methods of measuring term weight in the BOW representation. In table 2 we see that the precision and recall figures for TC dramatically improve when frequency is used as a means of measuring term weight. The accuracy on the test set increases considerably from 60% to 86.67%.

Because of the dramatic improvement in the performance of the SVM when frequency is used, we trained the SVM on each of the datasets and observed the effect in the figure below.

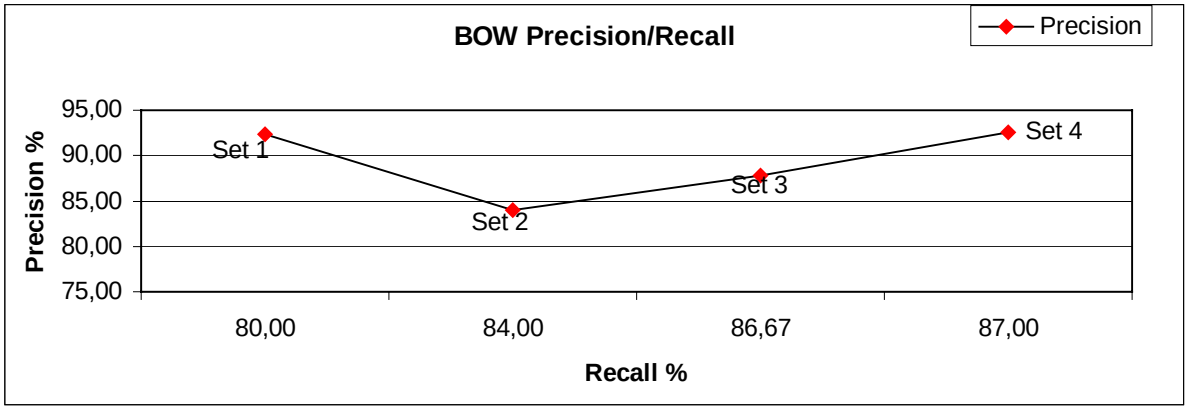


Figure 1. BOW precision and recall figures

Figure 1 illustrates how recall improves as the training set increases. Though precision figures took a drop when the training set was increased to 400, thereafter a steady increase was reported with the precision/recall figures for the final dataset achieving 92.55%/87.00%, a considerable improvement on precision/recall for set 1.

5.3.2. *N-gram word sequences*

In PRINCIP experiments, consistency analysis was performed on n-grams word sequences of length 2 and 3. Again certain n-grams (e.g. *our own kind, white civilisation, white survival, only Jews, our country)* were encountered significantly more often in the racist corpus showing that they are potentially discriminating.

Further SVM experiments will be carried out, first of all using the number of occurrences and then frequency as term weight. Both bigrams and trigrams will be investigated for each dataset. The results will then be compared to those obtained for BOW and POS.

5.3.3. *Parts-of-Speech*

The corpus was tagged using Xelda, a suite of linguistic tools made available to us by Xerox. The distribution of different POS across the three corpora was investigated.

	Racist	*Neutral*	*Anti-racist*
ADJ	8.89	8.58	7.7
ADV	4.61	3.7	3.29
NOUN	21.14	22.07	22.74
VERB	14.79	12.28	12.08
OTHER	50.57	53.38	54.19

Table 3. Distribution of the parts of speech in each corpus

The results (in table 3) shows there to be differences of between 1-3% across the board. These differences may seem insignificant and rather small but in order to put these figures into context the size of the samples must also be taken into consideration. It can be observed that the figures for the neutral corpus float in between the racist and anti-racist corpora. It is interesting to note that the racist corpus contains more adjectives and if we look at the adjective-noun ratio, (.42 for the racist and only .33 for the anti-racist) this tells us that racist discourse contains more qualifiers. The larger number of nouns in the anti-racist corpus may be indicative of a difference in register.

These differences, together with the results of Finn, Kushmerick and Smyth (Finn and Kushmerick, 2003; Finn *et al.*, 2002), are enough to justify training a SVM on POS datasets.

6. Conclusions and Future Research

In this paper we have described the field of text categorisation and its relationship to Information Retrieval and Machine Learning. We have introduced the various steps and processes involved in building a classifier and outlined the different choices to be made in doing so. We have introduced the PRINCIP problem and outlined how we are approaching the development of an automatic TC for racist texts on the Internet. We have presented our initial findings, evaluating the effectiveness of a SVM classifier trained on the bag-of-words representation. Though not 100% accurate we have shown it is possible to develop a TC system for racism on the web.

Our future research involves training Support Vector Machines for n-gram word sequences and POS – so as to identify the most effective method that will allow for the classification of racist documents on the web.

References

Apté C., Damerau F. and Weiss S.M. (1994). Automated Learning of Decision Rules for Text Categorization. *ACM Transactions on Information Systems.*

Chai K.M.A., Ng H.T. and Chieu H.L. (2002). Bayesian Online Classifiers for Text Classification and Filtering. In *Proceedings of the 25th annual international ACM SIGIR Conference on Research and Development in Information Retrieval:* 97-104.

Chandrasekar R. and Srinivas B. (1998). Using Syntactic Information in Document Filtering: A Comparative Study of Part-of-Speech Tagging and Supertagging. *Information Processing and Management,* vol. (34/5): 623-640.

Cohen W.W. and Singer Y. (1998). Context-sensitive Learning Methods for Text Categorization. In *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval*.

Cristianini N. and Shawe-Taylor J. (2000). *Support Vector Machines and other kernel-based learning methods.* Cambridge University Press.

Drucker H., Shaharary B. and Gibbon D.C. (2001). Relevance Feedback using Support Vector Machines. In *Proceedings of the 18th International Conference on Machine Learning*.

Finn A., Kushmerick N. and Smyth B. (2002). Genre classification and domain transfer for information filtering. In *Proceedings of the European Colloquium on Information Retrieval Research (Glasgow).*

Finn A. and Kushmerick N. (2003). Learning to classify documents according to genre. In *IJCAI-2003 Workshop on Computational Approaches to Text Style and Synthesis (Acapulco*).

Fürnkranz J. (1998). *A Study Using N-gram Features for Text Categorization.*

Internet Content Rating Association. http://www.icra.org/ Last visited 20/10/2003.

Gibbon M. and Greevy E. (2003). *The Truth About Racism*. Faculty of Humanities Research Seminar, Dublin City University, April 2nd 2003.

Goller C., Löning J., Will T. and Wolff W. (2000). Automatic Document Classification: A thorough Evaluation of various Methods. In *Proceedings of Der zweite Workshop des MK² zumThema Automatische Dokumentenklassifikation*. http://www11.informatik.tu-muenchen.de/forschung/foren/mkmk/proceedings/dokumenten/goller.pdf – [last visited 23/07/03]

Joachims T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the 10th European Conference on Machine Learning*.

Lechleiter H. and Greevy E. (2003). The Language of Open Racism: A Corpus Linguistic Analysis. In *Societas Linguistica Europea Conference*. Lyon, September 4th 2003.

Lewis D.D. (1992). Feature Selection and Feature Extraction for Text Categorization. In *Speech and Natural Language: Proceedings of a workshop:* 212-217.

Lewis D.D. and Ringuette M. (1994). A Comparison of Two Learning Algorithms for Text Categorization. In *Proceedings 3rd Annual Symposium on Document Analysis and Information Retrieval*.

Martin P. (2003a). So or Also: Racist use of adverbial phrases. In *Societas Linguistica Europea Conference.* Lyon, September 4th 2003.

Martin P. (2003b). Absolute Relatives – the language of online racial identity. In *Proceedings of the 30th Annual Symposium of the Royal Irish Academy*.

Mladenic D. and Grobelnik M. (1998). *Word sequences as feature in text-learning.*

Safer Internet Action Plan projects. http://www.saferinternet.org/filtering/projects.asp Last visited 20/10/2003.

Salton G. and Buckley C. (1988). Term-weighting Approaches in Automatic Text Retrieval. *Information Processing and Management,* vol. (24/5): 513-523.

Sebastiani F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, vol. (34/1): 1-47.

Smeaton A.F. (1997). Information Retrieval: Still Butting Heads with Natural Language Processing? In Pazienza M.T. (Ed.), *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*. *Computer Science*, vol. (1299). Springer-Verlag Lecture Notes: 115-138.

Smeaton A.F. and Kelledy F. (1997). Automatic Phrase Recognition and Extraction from Text. In Furner J. and Harper D.J. (Eds), *Proceedings of the 19th Annual BCS-IRSG on IR Research,* Aberdeen.Springer Electronic Workshops in Computing.

Tan C.M., Wang Y.F. and Lee C.D. (2002). The Use of Bigrams to Enhance Text Categorization. *Information Processing and Management,* vol. (38/4): 529-546.

Van Dijk T. (1987). *Communicating Racism. Ethnic Prejudice in Thought and Talk*. Newbury Park, CA Sage.Wilbur J. and Sirotkin K. (1992). The Automatic Identification of Stop Words. *Journal of Information Science*, vol. (18): 45-55.

Wodak R. and Reisigl M. (2001). *Discourse and Discrimination. Rhetorics of racism and anti-Semitism*. Routledge.

Xelda http://www.mkms.xerox.com/ Last visited 15/01/2004.

Yang Y. and Wilbur J. (1996). Using Corpus Statistics to Remove Redundant Words in Text Categorization. *Journal of the American Society Information Science (JASIS).*

Yang Y. (1999). An Evaluation of Statistical Approaches to Text Categorisation. *Journal of Information Retrieval*, vol. (1): 67-88.

Yang Y. and Liu X. (1999). A re-examination of text categorization methods. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*: 42-49.

Réflexions sur le traitement automatique des langues

Gaston Gross

LLI - UMR 7546 – Université Paris 13
Av. J-B Clément – 93430 Villetaneuse – France
gross@lli.univ-paris13.fr

The aim of this article is to examine two properties of natural languages which both constitute a hindrance to automatic processing and to propose solutions in each case. The first hindrance to any statistical processing of the vocabulary of a text, set phrases, is a constitutive feature of natural language and which occurs on a massive scale and which clearly complicates cutting a text up into units. The most natural way of solving this problem is to make an electronic dictionary with the same information given as for single words. The second difficulty is polysemy. All word categories are syntactically polysemous. It is therefore impossible to even start to analyse a sentence without being able to recognise meaning in context. For this, we put forward the idea of usage, which for a predicate means giving its argument pattern, which thereby defines meaning, actualisation (tense and mode). To describe argument patterns we put forward the idea of object classes, which is a perfectly natural way of factorizing analyses and taking inheritance into account.

Résumé

Cet article a pour objet d'examiner deux propriétés des langues naturelles qui constituent autant d'obstacles au traitement automatique et de proposer des solutions possibles dans chaque cas. Le premier obstacle à un traitement statistique du vocabulaire d'un texte est le problème du figement, qui est une des propriétés constitutives des langues naturelles et qui constitue un phénomène massif. C'est ici le découpage d'un texte en unités qui est en cause. La façon la plus naturelle de régler ce problème consiste à élaborer des dictionnaires électroniques portant les mêmes indications que pour les mots simples. La polysémie est une seconde difficulté. Toutes les catégories sont syntaxiquement polysémiques. Aussi est-il impossible d'envisager l'analyse d'une phrase sans être en mesure de reconnaître le sens en contexte. Nous proposons à cet effet, la notion *d'emploi* qui constitue, pour un prédicat, à donner son schéma d'arguments, qui définit à son tour le sens, son actualisation (temps et aspect). Pour la description des schémas d'arguments nous proposons la notion de *classes d'objets*, qui permet de façon très naturelle de factoriser les analyses et de rendre compte des héritages.

Abstract

Mots-clés : traitement automatique, figement, polysémie, classes d'objets, héritage.

1. Examen critique des conditions linguistiques du traitement statistique des textes

Le traitement statistique des textes pose en premier lieu le problème de la reconnaissance des unités lexicales constitutives. Compter le nombre de « mots » n'est pas une activité triviale. Ecartons d'abord comme peu problématique une particularité des langues naturelles qui influe sur le dénombrement des mots, à savoir l'amalgame : dans la suite *loi du moindre effort* faut-il considérer *du* comme un seul ou deux mots différents ? La syntaxe permet la plupart du temps de résoudre la problème. Deux propriétés des langues naturelles constituent en revanche un obstacle de taille à tout comptage : le figement et la polysémie. Voyons tout d'abord le figement. La définition du mot comme suite de caractères figurant entre deux blancs est une conception naïve du mot. Quel intérêt y aurait-il à décomposer *loi du moindre effort* en quatre unités constitutives et *preuve par neuf* en trois ? La question est d'autant plus importante que

20 à 30 % environ de la surface d'un texte (communication personnelle de M. Gross) est constituée par des mots complexes et que le nombre de mots composés est de l'ordre de plusieurs centaines de milliers. Les statistiques doivent donc porter non sur les éléments lexicaux formels mais sur les unités fonctionnelles, qui ont leur place dans les dictionnaires. Le système sera alors en mesure de les délimiter tout autant que les mots monolexicaux.

Mais, il y a plus important. Faire des statistiques sur les textes implique, par exemple, que l'on puisse déterminer si une forme comme *N de N* constitue une seule unité lexicale ou deux ou trois. La première décision que doit prendre un système en vue de l'analyse des groupes *N de N* consiste à déterminer si les suites ainsi formées sont des libres ou figées. Comme ces dernières doivent figurer dans le dictionnaire (elles ne peuvent pas faire l'objet d'un calcul, leur sens n'étant pas compositionnel), leur reconnaissance tient lieu d'analyse. C'est le cas, par exemple, des noms composés du type *pomme de terre, tour de vis, tête de pont.* Mais il faut dans ce domaine prendre en compte la suite la plus longue du figement. Il existe un composé *corps de garde* que le système doit reconnaître et qui est codé dans le dictionnaire à la fois comme un humain et un locatif. Mais on doit aussi être en mesure de regarder l'environnement immédiat pour éviter de disloquer un nom composé plus long comme *plaisanterie de corps de garde*, dont il est un élément constitutif. De telles analyses ne peuvent se faire que sur la base d'un dictionnaire indiquant les degrés de figement et les sous-ensembles figés de suites (figées) plus longues. Cette précaution permettra de reconnaître *station de sports d'hiver, bande d'arrêt d'urgence, huile de foie de morue, excédent de la balance des paiements, règle de non-affectation des recettes, chemin de fer de ceinture.*

2. La polysémie

Le problème posé dans la section précédente est connu depuis longtemps de ceux qui s'intéressent au traitement automatique des langues, même si l'importance fondamentale des dictionnaires dans le traitement automatique n'est pas partagée par tous. La polysémie constitue une autre difficulté théorique au traitement statistique des textes et pourrait représenter une forte limitation à l'intérêt que présente ce type de recherches. Dès les premiers travaux de statistiques, il était de tradition de séparer les formes lexicales appartenant à des catégories différentes (*le* article et pronom ; *que* relatif ou conjonctif).

Mais des distinctions doivent être établies aussi à l'intérieur des catégories elles-mêmes. Le comptage des verbes est une opération assez simple puisque leur morphologie spécifique permet leur reconnaissance, exception faite des formes de la troisième personne du singulier de l'indicatif qui sont souvent homographes des formes nominales. Mais même les dictionnaires actuels les plus détaillés, comme par exemple Bescherelle, ne sont pas entièrement satisfaisants. Dans ces ouvrages, les verbes sont regroupés selon leur conjugaison. C'est oublier qu'une même forme infinitive peut appartenir à deux types de conjugaisons différents en fonction de ses emplois. Le verbe *pleuvoir* appartient à la table 47 de Bescherelle, qui recense les verbes défectifs n'ayant que le 3^{e} personne du singulier : *il pleut il pleuvait, il pleuvra*. Mais il existe un autre verbe *pleuvoir*, qui est un verbe support d'occurrence et que l'on trouve dans des phrases comme *Les coups pleuvaient sur la tête de Paul*. On voit que si on n'est pas en mesure de séparer les emplois, l'analyse n'a guère d'intérêt puisqu'on ne voit rien de commun à ces deux emplois. Donc, pas de morphologie sans syntaxe.

De façon plus générale, l'étude de la fréquence des verbes d'un texte ne peut se faire que si on est en mesure de reconnaître le statut syntaxique de chaque forme verbale. Or, il existe au moins six types de verbes différents.

a) Les plus fréquents sont évidemment les verbes prédicatifs, qui ont des arguments et que la tradition assimile à la catégorie.

b) Les verbes figurant dans les expressions verbales figées et qui ne doivent pas être confondus avec les précédents, dans la mesure où on ne peut pas parler d'arguments à leur propos : dans *prendre la poudre d'escampette*, le substantif ne peut pas être assimilé à un argument du verbe *prendre*. Pour éviter de leur attribuer des propriétés distributionnelles, il faut là encore les lister et les traiter comme des blocs inanalysables qui doivent être traduits comme tels.

c) Les pro-verbes qui fonctionnent comme des substituts de classes de verbes assez générales. C'est le cas de *faire* qui sert d'anaphore aux prédicats d'action.

d) L'emploi causatif de *faire (faire travailler les enfants)*, qu'on ne doit pas confondre avec le précédent. Voici quelques autres causatifs : *provoquer (un incendie), rendre (rendre fou), mettre (mettre en difficulté)*. Les causatifs sont des verbes qui opèrent sur d'autres prédicats. Ils figurent donc dans des phrases complexes.

e) Les deux derniers types de verbes représentent des auxiliaires. Les plus connus sont les auxiliaires verbaux, qui traduisent le temps mais plus souvent l'aspect (*aller, venir de, être sur le point de*).

f) Enfin, il existe des auxiliaires de prédicats nominaux, les verbes supports, qui conjuguent ces prédicats. Ces verbes sont donc syntaxiquement très différents, ils ne sont jamais prédicatifs.

Le simple décompte des formes verbales dans un texte ne rendrait pas compte de ces différences et n'apporterait aucune information linguistique pertinente.

3. Propositions : la notion d'emploi

Les observations que nous venons de faire montrent que le travail de description ne peut pas prendre pour argent comptant la notion de catégories grammaticales, car toutes peuvent être syntaxiquement polysémiques. Mais on doit aller plus loin. Si on examine les verbes prédicatifs eux-mêmes, force est de constater que tous sont polysémiques et qu'il est donc illusoire par exemple de parler dans un texte « du » verbe *tenir*. Une rapide description de ce verbe permet de montrer un grand nombre d'emplois différents, de sorte que des statistiques portant sur la forme morphologique elle-même n'ont qu'un intérêt réduit. L'idée d'un classement sémantique des lexèmes a été envisagée par Ch. Muller (Préface à Lafon, 1984) « (Le classement sémantique) est une question de temps et de moyens plus que de théorie ; on ne voit guère quelle objection de principe pourrait être opposée à des pratiques aussi obligeantes ». Abordons l'étude de *tenir* par les emplois de supports, c'est-à-dire d'actualisateurs de prédicats nominaux :

Paul tient un discours politique

Paul tient une bonne cuite

Il y a des emplois d'opérateurs à lien :

Cet enfant est sous l'autorité du juge

Le juge tient cet enfant sous son autorité

Parmi les emplois prédicatifs, on peut distinguer d'abord les emplois de type « locatif » :

Paul tient cet enfant par la main

Paul tient ce bijou dans sa main

Ce bagage ne tient pas dans cette malle

Cela tiendrait dans le creux de ma main

D'autres sont appropriés à certaines activités commerciales :

Paul tient un commerce de voitures

Paul tient un stand au marché de Noël

Il y a un emploi « passif » de prédicats :

de <don> ou de <transmission> *:*

Paul tient cette nouvelle de son voisin

Paul tient cette maison de son grand-père

de sentiments :

Paul tient à ce voyage

de « ressemblance » :

Cet enfant tient de sa mère

Ce discours tient du délire

aspectuels :

La promesse tient

La colle tient

Le pont tient

Ces emplois sont si différents que toute manipulation qui les assimilerait ne rendrait pas compte de leur fonctionnement linguistique. Il faut donc être en mesure de reconnaître à quel ensemble appartient un mot, c'est-à-dire de préciser son emploi. Quand nous parlons d'emploi, nous entendons plusieurs paramètres intégrés. Un emploi prédicatif est défini par :

a) Un domaine d'arguments (ces arguments, dont on note la suite la plus longue, sont définis à l'aide des classes d'objets) ;

b) Un sens, d'où par conséquent un ou plusieurs synonymes, un antonyme et une traduction ;

c) Une forme morphologique : *verbe*, *nom*, *adjectif* pour les prédicats des phrases simples (prédicats du premier ordre) ; *prépositions* et *conjonctions* pour les prédicats du second ordre ;

d) Une actualisation : conjugaison pour les verbes, verbes supports pour les prédicats nominaux. Certains prédicats sont défectifs du point de vue de leur actualisation : *regarder* (concerner) n'a pas de passé composé ;

e) Un système aspectuel. Il doit y avoir compatibilité aspectuelle entre les différents éléments porteurs d'aspect dans la phrase (déterminants des arguments, adverbes, adjectifs, etc.). *Peur* et *peureux* n'ont pas le même aspect et ne peuvent pas être considérés comme constituant le même prédicat : *peur* est un sentiment qui peut être ponctuel ; *peureux* ne désigne pas un sentiment mais un trait de caractère et est nécessairement duratif. En revanche, *désirer* et *désireux* ont le même système aspectuel ;

f) Des restructurations (transformations) qui lui sont propres : passivation, thématisations différentes, pronominalisations, etc.

L'information la plus importante est constituée par le schéma d'arguments, car c'est de lui que dépendent les autres propriétés. Or, pour mettre en évidence ces schémas, il faut être en mesure de décrire les arguments avec précision.

4. Les classes d'objets

C'est un fait admis qu'un prédicat est d'abord défini par son domaine d'arguments, c'est-à-dire son sujet et ses compléments. Nous avons vu que la plupart des prédicats sont polysémiques. C'est une observation empirique que tout changement de sens d'un prédicat est corrélé à un changement de son schéma d'arguments. Soit la phrase *Vous suivrez ce chemin*. Si on remplace l'objet *chemin* par des substantifs comme *route, rue, voie, sentier* le verbe *suivre* garde le même sens. On regroupera ces mots sous le terme générique de <voies>. Si en revanche, on remplace le mot *chemin* par *cours*, alors on a affaire à un autre emploi et le substantif *cours* peut être remplacé par *séminaire, stage, formation, cycle d'étude*, etc., qu'on rangera sous le classifieur <enseignement>. Le sens du verbe *suivre* serait encore différent si le complément était *recommandation, suggestion, avis* qu'on classerait comme <conseil>, ou encore *cure, médication, régime , thérapeutique* qui relèverait de la classe des <traitements>. La mise au point du sens exige que l'on soit à même de préciser la nature sémantique des arguments que prend un emploi donné de prédicat. Les ensembles lexicaux représentant les arguments en compréhension s'appellent *classes d'objets*. Il est donc indispensable qu'une information comme <voies >, <enseignement>, <conseil >, <traitement> correspondant à des *classes d'objets*, figure dans le dictionnaire comme classifieurs de substantifs décrivant des positions argumentales.

5. Opérations sur la base de ces descriptions

Tous les substantifs figurant dans une même position argumentale pour un sens déterminé appartiennent à la même classe d'objets. Dès lors que l'on a séparé les différents emplois d'un prédicat à l'aide des classes d'objets, on est en mesure de reconnaître ou de générer toutes les phrases correspondant à chacun des emplois. Le sens d'un prédicat est donc fonction du schéma d'arguments. Ainsi, le verbe *régler* peut avoir comme arguments *régler/N0 :hum/N1 :<facture>/N2 : à hum*. Appartiennent à la classe <*facture*> les éléments suivants : *addition, compte, état de frais, facture, note, relevé de compte*. À cette liste, il faut ajouter des termes populaires (*douloureuse*) ainsi qu'un très grand nombre de noms composés, où le complément en *de N* spécifie l'objet du règlement : *note d'électricité, note d'honoraires, note d'hôtel, note d'un artisan, note d'un entrepreneur, note de blanchisseuse, note de crédit, note de droit d'auteur, note de frais de transport, note de frais, note de gaz, note de manucure, note de pressing, note de restaurant, note de téléphone***.**

L'ensemble de ces mots forme une des classes sémantiques possibles en position d'objet du verbe *régler*. Cette classe ne doit pas être confondue avec une autre qui désigne un reçu représentant une attestation de paiement : *attestation, bon de caisse, bulletin de bagages, bulletin de consigne, justificatif, récépissé, reçu, vignette, reçu de carte bancaire, ticket de caisse*.

Inversement, on peut aussi se demander quels sont, pour une classe donnée, les opérateurs qui lui sont appropriés. Dans la démarche homographique que nous adoptons et qui considère que les différents emplois d'un prédicat polysémique constituent des ensembles disjoints, un argument n'est pas essentiellement défini par ses traits sémantiques inhérents mais par l'ensemble des prédicats qui lui sont strictement appropriés.

acquitter/N0:<acheteur>/N1:<facture>/N2:

annuler/N0:<acheteur>/N1:<facture>/N2:

augmenter/N0:<vendeur>/N1:<facture>
baisser/N0:<vendeur>/N1:<facture>/N2:
corser/N0:<vendeur>/N1:<facture>/N2:
demander/N0:<acheteur> /N1:<facture>/N2:à <vendeur>
dresser/N0:<vendeur>/N1:<facture>/N2: à l'ordre de <acheteur>
établir/N0:<vendeur>/N1:<facture>/N2: à l'ordre de <acheteur>
fournir/N0:<vendeur>/N1:<facture>/N2:à <acheteur>
grossir/N0:<vendeur>/N1:<facture>/N2:
honorer/N0:<acheteur>/N1:<facture>
payer/N0:<acheteur>/N1:<facture>/N2:à <vendeur>
présenter/N0:<vendeur>/N1:<facture>/N2:à <acheteur>
recevoir/N0:<acheteur>/N1:<facture>/N2: de <vendeur>
rédiger/N0:<vendeur>/N1:<facture>/N2:
réduire/N0:<vendeur>/N1:<facture>/N2: de < %>
régler/N0:<acheteur>/N1:<facture>/N2:à <vendeur>
s'élever/N0:<facture>/N1: à Card<unité monétaire>
solder/N0:<acheteur>/N1:<facture>/N2:

Les prédicats appropriés de *<reçu>* sont entre autres les suivants :

antidater/N0:<vendeur>/N1:<reçu>/N2:
dater/N0:<vendeur>/N1:<reçu>/N2:
délivrer/N0:<vendeur>/N1:<reçu>/N2:à <acheteur>
dupliquer/N0:<vendeur>/N1:<reçu>/N2:
exiger/N0:<acheteur>/N1:<reçu>/N2: de <vendeur>
fournir/N0:<vendeur>/N1:<reçu>/N2: à <acheteur>
recevoir/N0:<acheteur>/N1:<reçu>/N2: de <vendeur>
remettre/N0:<vendeur>/N1:<reçu>/N2:à <acheteur>
valider/N0:<vendeur>/N1:<reçu>/N2:
viser/N0:<vendeur>/N1:<reçu>/N2:

Au regard des verbes que nous venons d'énumérer les substantifs de la classe des <factures> se comportent de la même façon. Les distinctions qu'on peut établir ne sont pas de nature syntaxique mais pragmatique. Une *addition* est une <facture> que l'on établit dans un restaurant, une *note* est la <facture> qu'on paie dans un hôtel, un *état de frais* est établi par un employé à l'intention de son employeur, un *relevé (de compte)* représente la <facture> que l'on reçoit d'une administration prestatrice de services (EDF, GDF). Ces informations peuvent être ajoutées dans une base de données en ouvrant un champ indiquant les domaines.

6. Classes linguistiques et classes référentielles

La notion de classes (et d'hyperclasses) a comme premier avantage de pouvoir factoriser et par là de décrire de façon compacte tous les éléments d'une classe d'objets donnée. Ainsi tous les éléments d'une classe héritent de l'ensemble des prédicats appropriés de celle-ci, comme nous venons de le voir. Nous abordons ici la notion générale d'héritage et nous commençons par des cas où les ensembles sont homogènes. Le remplacement des éléments par leur classe

représente une grande simplification de la description. Si d'une part, on attribue dans un dictionnaire électronique à chaque substantif le code de la classe à laquelle il appartient et si, d'autre part, on décrit les schémas d'arguments des prédicats à l'aide de ces classes, alors on est en mesure de reconnaître ou de générer l'ensemble des phrases appartenant à un emploi donné. Mais cela n'est possible que si l'on a pris soin de créer des classes « linguistiques », c'est-à-dire des classes qui comprennent des substantifs ayant exactement les mêmes propriétés sémantiques et syntaxiques. Dans notre démarche, un mot est défini par son environnement et non en lui-même, comme c'était le cas dans l'analyse sémique traditionnelle. Nos classes représentent donc des ensembles lexicaux, l'ensemble des <factures>, l'ensemble des <habitations>, l'ensemble des <unités monétaires>.

Mais tous les ensembles ne constituent pas des classes au sens où nous l'entendons ici. Prenons un terme collectif comme les *effets personnels*, défini ainsi par le Nouveau Grand Robert « 3. (XVII^e^). *Cour. Le linge et les vêtements. – Affaire (affaires), défroque, fringue, frusque, habit, harde, nippe, trousseau, vêtement. Mettre ses effets dans une valise. Ballot d'effets. – Bagage. Les effets d'un militaire. – Paquetage. Effets civils, militaires.* ». On voit que ce terme désigne des éléments qui n'ont pas le même comportement syntaxique : les prédicats appropriés au mot <linge> ne sont pas ceux qui s'appliquent aux <vêtements> : on met un vêtement mais non un linge. Il n'est pas clair non plus si le terme « bagage » est compris dans la définition. On a affaire ici à des classes référentielles mais non linguistiques.

La même observation peut être faite avec un terme comme *vaisselle* que le NGR définit « 2. (XIX^e^). Ensemble des plats, assiettes, ustensiles de table, etc., qui sont à laver. *Laver, faire la vaisselle. – Nettoyer, relaver* (régional). Laver les vaisselles. Écurer (vx), égoutter, rincer, essuyer la vaisselle. Laveur de vaisselle. – Plongeur. Machine à laver la vaisselle. – Lave-vaisselle. Laisser s'entasser la vaisselle. – Bac à vaisselle ». Il est clair que les contenants (*plat, assiette, saucier*, etc.) n'ont pas les mêmes opérateurs appropriés que les ustensiles de table (*cuiller, fourchette, couteau*). Ces derniers ne constituent pas non plus une classe « linguistique ». Il va de soi, d'autre part, que les classes doivent pouvoir faire l'objet d'une énumération objective et donc d'un consensus. Un ensemble comme les « choses écœurantes » ne constitue pas une classe linguistique, car une description en extension serait aléatoire.

Prenons un autre exemple. Les <sports> ont comme opérateur approprié le verbe *pratiquer*, les <mouvements> *effectuer*, les <matières scolaires> *faire (du)*. Il serait raisonnable de penser que ces termes ont comme hyperonyme le mot *activité*, dont le verbe approprié est *exercer : il exerce une activité débordante*. Or, les hyponymes n'héritent pas cet opérateur **exercer du foot, *exercer une promenade, *exercer le latin.* On ne classera donc pas ces diverses actions sous le terme générique d'activité.

7. Classes et sous-classes

Nous avons vu que les substantifs de la classe des <factures> ont tous le même comportement syntaxique : ils prennent les mêmes prédicats appropriés. Il n'y a donc aucune raison de les sous-catégoriser du point de vue linguistique. Les différences observées relèvent de considérations pragmatiques, comme nous l'avons vu : telle facture est propre aux restaurants (*addition*) , telle autre aux hôtels (*note*) ou encore aux relations professionnelles (*état de frais*). Mais la description adéquate d'ensembles lexicaux nécessite la plupart du temps que l'on subdivise les classes en sous-ensembles. Prenons l'exemple de la classe des <boissons>. Observons d'abord que les notions de *boire* et de *boisson* ne sont pas à mettre sur le même plan. Au sens strict du mot, *boire* a comme objet un <liquide>, car on peut boire par inadver-

tance des liquides non destinés à cet effet : produits pharmaceutiques, carburants, etc. Beaucoup de produits d'entretien portent l'indication « ne pas avaler ». Le terme de <boisson> ne doit donc pas être confondu avec les compléments possibles du verbe en position d'objet. Une boisson est un <liquide> destiné à être bu. Cette classe exclut les liquides dont nous avons parlé. Cela dit, on peut alors faire le recensement de tous les prédicats appropriés à cette classe dans son ensemble :

Verbes : *boire, siroter, siffler*

Sont communs aux aliments : *absorber, avaler, ingurgiter*

Prédicat nominal (aspect itératif) *: être buveur de*

Adjectifs : *froid, glacé, tiède, chaud, brûlant, doux, insipide, fade, buvable, imbuvable, potable, non potable*

Les opérateurs que nous venons de donner sont communs à toutes les boissons. Il faut ensuite mettre au point des sous-classes. Une première grande division est celle qui sépare les boissons alcoolisées des autres. On procédera pour les boisson alcoolisées à la même factorisation en mettant en évidence les opérateurs généraux puis en se servant de nouveaux opérateurs pour créer des sous-classes. Parmi les opérateurs appropriés aux <boissons alcoolisées> on trouve en position de sujet, entre autres :

*titrer /N0:<alcool>/N1 :[*Card] *degré(s),*
enivrer/N0:<alcool>/N1:hum,
saouler/N0:<alcool>/N1:hum

En position d'objet avec un sujet humain, on peut relever : *cuver, frelater, picoler, pinter, pitancher, trafiquer*. Les prédicats nominaux sont entre autres : *avoir* [Card] *degré(s), avoir une teneur en alcool de* [Card] *degrés*. Et les prédicats adjectivaux : *âpre, doux, léger, lourd, moelleux, raide, sec.*

À l'aide d'autres opérateurs on peut établir des sous-classes. Les <alcools et spiritueux> prennent le verbe *distiller (N0:hum/N1:<alcools et spiritueux)*. Les prédicats appropriés aux <vins> sont plus nombreux :

Verbes :

accompagner/N0: <vin>/N1 :<plat>
débourber/N0:hum/N1: <vin>
décanter/N0:hum/N1: <vin>
se madériser/N0: <vin>
tirer/N0: <vin>

Adjectifs :

aigre/N0:<vin>
aigrelet/N0:<vin>
âpre/N0:<vin>
astringent/N0:<vin>
bourru/N0:<vin>
capiteux/N0:<vin>
charnu/N0:<vin>
charpenté/N0:<vin>
corsé/N0:<vin>
équilibré/N0:<vin>
gouleyant/N0:<vin>

grêle/N0:<vin>
harmonieux/N0:<vin>
pétillant/N0:<vin>
piquant
piqué/N0:<vin>
tuilé/N0:<vin>
vert/N0:<vin>
vieux/N0:<vin>

Un verbe comme *brasser* s'applique aux <bières>, qui prennent aussi des adjectifs comme :

aigre/N0:<bière>
amère/N0:<bière>
filante/N0:<bière>
forte/N0:<bière>
plate/N0:<bière>

8. Relations d'héritage

Un élément lexical donné prend, bien entendu, les opérateurs qui sont strictement appropriés à sa classe. Ainsi, un <vin> est défini par des prédicats verbaux comme *madériser* ou adjectivaux comme *gouleyant* ou *charnu*. Mais comme relevant de l'hyperclasse <alcool>, il hérite de l'ensemble des opérateurs qui définissent ce niveau : il peut *enivrer, monter à la tête, saouler* ; on peut le *cuver*, le *tenir*. Il peut avoir une *teneur de* [Card] *degrés*. Au niveau supérieur, comme tout alcool est une boisson, le vin peut *se boire*, s'*absorber*, *se siroter*, *se siffler* par un *buveur de vin*. Il peut être *froid, glacé, tiède, chaud, doux, insipide, fade, buvable, imbuvable, potable, non potable*. Mais on peut aussi subdiviser les vins à l'aide des adjectifs *rouge, blanc, rosé, vert, jaune*. Ces adjectifs ne sont pas des qualificatifs mais des désignatifs caractérisant différents types de vins. Cette sous-classification est référentielle mais non linguistique, car elle ne permet pas de mettre en évidence des opérateurs qui seraient appropriés à chacun de ces types de vins. Les vins peuvent encore être subdivisés en appellations et noms de marques. Là non plus, nous n'avons pas affaire à de vraies sous-classes linguistiques, car elles ne génèrent pas non plus de syntaxe spécifique.

Si l'on devait faire une arborescence rendant compte de la syntaxe des noms de <vins>, on partirait au niveau la plus élevé de la notion de <concret>. De ce fait, ces substantifs hériteraient de toutes les propriétés générales des concrets : poids, volume, couleur, etc. Ensuite on peut hésiter sur l'ordre de deux traits : <artefact> ou <liquide>. Nous choisissons d'abord le trait <liquide>, ce qui permet de prédire des verbes comme : *verser, couler, déborder, imbiber, s'égoutter* ou des adjectifs comme *dense, fluide, huileux*. Ensuite, une subdivision séparera les liquides naturels (dont on vient de voir la syntaxe) des <(liquides) artefacts>, qui ont des verbes comme *fabriquer, réaliser, mettre au point* mais aussi *vendre, acheter, avoir tel ou tel prix* et tous les autres prédicats pouvant caractériser les produits commerciaux. Ensuite, on notera qu'il s'agit de <boissons>, puis de <boissons alcoolisées> et enfin de <vins> et on aura les opérateurs appropriés que nous avons notés plus haut. Un tel travail descriptif est d'abord un problème linguistique plus que de représentation informatique.

9. Héritage et métaphore

La définition du sens à l'aide de l'environnement permet de détecter des emplois métaphoriques. Prenons les moyens de transports. En tant que tels, ils ont des opérateurs généraux *se déplacer en, voyager en, aller en, arrêter, descendre de, monter en, prendre,* etc. Pour les décrire, il y a deux grands paramètres d'analyse. Le premier met en jeu le mode de transport : terrestre, maritime, ou aérien. Chacun de ces types de transports a des prédicats appropriés et dont la liste n'est pas difficile à établir : pour les avions : *atterrir, décoller, descendre en piqué, descendre en spirale, descendre en vrille, piquer, plafonner, planer, s'écraser, s'écraser au sol, se cabrer, se poser ;* pour les bateaux *appareiller, chavirer, démâter, dériver, faire*

eau, gîter, lever l'ancre, mouiller, s'échouer, tanguer ; pour les transports terrestres *caler, circuler, freiner, rouler, prendre la route de, verser, dépasser*.

À cela s'ajoute la distinction entre moyens de transports individuels et collectifs. Cette opposition est linguistique et pas seulement pragmatique. S'il est possible de *prendre* tout type de moyens de transports, *emprunter* n'est possible qu'avec les transports en commun. Ces derniers ont aussi comme particularité d'être suivis d'un horaire (*le train de midi, *la voiture de midi)* et d'une destination (*le train de Paris, *la voiture de Paris*). Ils ont aussi un grand nombre de prédicats appropriés. En position de sujet on trouve *desservir la destination de, accuser un retard de, annoncer un retard de, être à destination de, partir à n heures* et en position d'objets *attraper le dernier, embarquer dans, embarquer sur, emprunter, manquer son, louper, partir par, prendre le dernier, rater*.

Parmi les transports routiers, on peut isoler une sous-classe particulière, celle des <transports par animal> : *cheval, mulet, chameau, âne,* etc. On trouve alors des opérateurs appropriés : *voyager à dos de, faire une promenade à, monter N en amazone, monter, faire du, être à califourchon sur, se déplacer à dos de, tomber de, faire une chute de*. Il va de soi que ces animaux ne sont interprétés comme des moyens de transports qu'avec les opérateurs que nous avons mentionnés. D'autres verbes les feraient appartenir à la classe des animaux de traits : *brider, harnacher, bouchonner, ferrer, soigner, seller*.

Si maintenant on analyse le comportement des moyens de transports individuels appelés <deux-roues>, on observe qu'ils ont des opérateurs appropriés communs avec les <moyens de transports animaux>. À la différence des autres moyens de transports qui prennent la préposition *en (en bateau, en voiture, en train),* on a ici la préposition *à (être, aller, monter, faire un tour) à (cheval, vélo, moto*). On trouve aussi la préposition *sur : être perché sur (son vélo, son cheval)*. Ce sont des compléments naturels du verbe *enfourcher : enfourcher (son cheval, son vélo*). De plus, ils ont en commun des prédicats de mouvement : (*tomber, faire une chute) de (cheval, vélo, moto*). Observons encore qu'à la différence des autres moyens de transports, un <deux-roues> ne peut ni *partir* ni *arriver*. La métaphore est donc une particularité des langues naturelles qui interdit que l'on établisse des arborescences en dehors de la syntaxe.

10. Les unités lexicales complexes : héritages multiples ou autonomie ?

Nous avons vu avec les moyens de transports animaux que certains substantifs peuvent appartenir à plusieurs classes. Par exemple, un cheval appartient à la classe des <équidés> et par là des <mammifères>, on aura alors des opérateurs comme *pouliner, mettre bas*. Il peut aussi appartenir à l'ensemble des <animaux de traits> avec des verbes comme *atteler, dételer, harnacher* ou encore des <animaux de course> et on aura les verbes *monter, entraîner, jouer sur, miser sur*. Cela pose de façon générale le problème des arborescences et des conditions nécessaires à la création de nouvelles classes. Les substantifs qui sont compléments à la fois des verbes *porter, mettre, enfiler* et *ôter* sont des vêtements. Un substantif comme *cotte de maille* entre dans cette classe. Or, ce substantif peut être sujet d'un verbe comme *rouiller*. Faut-il de ce fait créer une nouvelle classe de <vêtements> ? La réponse est d'ordre statistique. On peut penser que les occurrences du verbe *rouiller* sont si rares qu'on peut le négliger. Mais cette remarque doit être étayée. Tout d'abord il existe d'autres noms de vêtements qui peuvent être en métal : *cuirasse, heaume, casque,* etc. D'autre part, d'autres verbes pourraient être appropriés à des objets métalliques *tinter, faire du bruit, résonner,* etc. Il y aurait donc intérêt à ouvrir une sous-classe de <vêtements de chevalerie>. D'autres cas vont en sens inverse. Le verbe *chausser* a comme compléments des substantifs de la classe des <chaussures> : *pantoufle, soulier, botte, espadrille, basket,* etc. Mais il existe un emploi où le verbe a

comme objet le mot *lunettes* (ou *bésicles*). Ces mots appartiennent à la classe des <prothèses> qui ont avec les <vêtements> beaucoup de verbes en commun : *porter, mettre, ôter* mais non *enfiler*. Faut-il créer une classe autonome <lunettes> du seul fait que l'on peut utiliser le verbe *chausser*, on peut hésiter. Cela dépend de l'objectif qu'on se fixe.

Dans d'autres cas, le problème se pose de façon plus sérieuse. La syntaxe du mot *livre* a été souvent examinée (cf. Kayser, 1987 et 1989 ; Kleiber et Riegel, 1989 ; Pujstejovsky, 1995 ; Pustejovsky et Bouillon, 1995). Comme nous définissons les arguments (i.e. les substantifs) par leurs prédicats appropriés, il est clair qu'un même mot entrera dans autant de classes qu'il sera défini par des séries prédicatives différentes. Cela est évident pour les homographes ou les mots polysémiques : chien (canidé : *aboyer*), chien (injure : *traiter qq de*), chien (pièce coudée de certaines armes à feu : *abattre*). Voyons le cas du mot *livre*. Il est caractérisé par des séries prédicatives très diverses. Il peut être interprété comme :

a) un concret : tenir (dans ses mains), peser (tant), être adj de couleur, tomber

b) un abstrait : être obscur, difficile, indéchiffrable

c) un humain : prétendre, affirmer, exposer, révéler

d) un locatif : contenir (n chapitres), comprend (des erreurs), parcourir (un livre)

e) un événement : *paraître, sortir*

Ici, il est difficile de dire que l'une des séries est plus fréquente ou plus naturelle ou disponible que les autres. Théoriquement, les substantifs qui relèvent de cette classe pourraient avoir des héritages multiples. Le problème posé ici n'est pas de savoir comment on peut représenter informatiquement les héritages multiples mais comment on doit rendre compte des faits linguistiques. Le mot *livre* doit être décrit de façon plus précise. Il est clair qu'un livre représente un <texte> et que de ce fait il est compatible avec des adjectifs comme *obscur, long, indéchiffrable* et des adjectifs comme *écrire* et *lire*. Mais un livre est aussi un <support de textes> comme *journal, revue, périodique*. Rappelons d'abord qu'on ne doit pas confondre les <supports de textes> avec les <supports d'écriture> : *cahier, ardoise, calepin*. Il y a une différence entre ces deux classes : on peut lire un livre ou un journal mais non un cahier ou une ardoise.

Dans l'interprétation du mot *livre* on est en présence d'une série de métonymies. Un texte peut par métonymie être assimilé à un humain, d'où des adjectifs communs : *obscur, incompréhensible (texte, auteur)*, mais *indéchiffrable* ne semble s'appliquer qu'à des textes. Le constitution d'un livre en chapitres en fait métaphoriquement un lieu. L'ensemble des prédicats que nous venons de donner ne s'appliquent qu'à la classe des <livres>. Il y a donc intérêt à considérer des classes de ce type comme autonomes et sans lien avec des hyperclasses, en quelque sorte comme des entités « autocéphales ». Cela pourrait multiplier les classes mais aurait l'avantage de la précision.

Conclusion

Dans ces pages, nous avons essayé de soulever quelques difficultés que présentent les langues au traitement automatique et en particulier aux lemmatiseurs. Nous nous sommes placés dans la perspective de la reconnaissance automatique et nous avons montré que chaque forme appartient à un emploi, au sens technique du mot que nous avons décrit au paragraphe 3. Cette notion d'emploi insère les éléments lexicaux dans des phrases où les schémas d'arguments spécifient le sens en contexte des prédicats. D'autres informations spécifiques à chaque emploi sont notées, de telle façon que le levée de la polysémie et en général des ambiguïtés

soit possible sur la base des propriétés, qui figurent dans un lexique électronique. Celles-ci peuvent être considérées dans les textes comme des indices permettant de reconnaître l'emploi effectif parmi un grand nombre d'autres possibles. Notre visée est donc celle d'une automatisation des procédures de reconnaissance.

Bibliographie

Bescherelle. (1997). *La conjugaison pour tous*. Hatier.

Gross G. (1989). *Les constructions converses du français*. Droz.

Gross G. et Clas A. (1997). Synonymie, polysémie et classes d'objets. *Meta*, vol. (42/1). Presses de l'Université de Montréal : 147-155.

Gross G. (1998). Pour une véritable fonction *Synonymie* dans un traitement de texte. *Langages*, vol. (131). Larousse : 103-114.

Gross G. (1999). Élaboration d'un dictionnaire électronique. *Bulletin de la Société de Linguistique de Paris*, Tome (XCIV/1). Peeters : 113-138.

Gross G. (1997). Les classes d'objets et le désambiguïsation des synonymes. *Cahiers de Lexicologie*, vol. (70). Didier Erudition : 27-40.

Gross G. (1998). Pour une typologie des prédicats. *Prédication, assertion, information*. *Studia Romanica Upsaliensis*, vol. (56). Uppsala : 221-230.

Gross G. (1999). La notion d'emploi dans le traitement automatique. *La pensée et la langue*. *Wydawnictowo Naukowe AP :* 24-35.

Gross G. et Guenthner Fr. (1999). Traitement automatique des domaines. *Revue Française de Linguistique Appliquée*, vol. (III/2) : 47-56.

Kayser D. (1987). Une sémantique qui n'a pas de sens. *Langages*, vol. (87) : 33-45.

Kayser D. (1989). Réponse à Kleiber et Riegel. *Lingvisticae Investigationes*, vol. (XIII/2) : 419-422.

Kleiber G. (1984). Polysémie et référence : la polysémie, un phénomène pragmatique ? *Cahiers de lexicologie*, vol. (44/1) :85-103.

Kleiber G. et Riegel M. (1989). Une sémantique qui n'a pas de sens n'a pas de sens. *Lingvisticae Investigationes*, vol. (XIII/2) : 405-417.

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Slatkine-Champion.

Pustejovsy J. et Bouillon P. (1995). Aspectual Coercion and Logical Polysemy. *Journal of Semantics*, vol. (12) : 133-162.

Pustejovsky J. (1995). *The Generative Lexicon*. The MIT Press.

Tournier M. (1985). Sur quoi pouvons-nous compter ? Réponse à Charles Muller. *Verbum, Hommage à Hélène Nais* : 481-492.

Nouvelle méthode d'analyse statistique d'apparition d'un mot particulier (études synchroniques et diachroniques)

Patricia Guilpin[1], Christian Guilpin[2]

[1] SYLED – CLA[2]T – ILPGA – Université de la Sorbonne Nouvelle Paris 3 – 19, rue des Bernardins – 75005 Paris – France

[2] Groupe de Physique des Solides – Universités Paris 6 et 7 – UMR 75 88 Campus Boucicaut 140, rue de Lourmel – 75015 Paris – France

guilpin@gps.jussieu.fr

Abstract

In this paper, a new statistical method is presented to study the frequency of a particular word at different times or in different styles (synchronic and diachronic studies). It is made use of equations one can easily compute. The example developed in order to validate our method deals with variations in the use of some determiners in classical Greek (by Aristophanes and Herodotus).

Résumé

Dans ce papier, nous développons une nouvelle méthode statistique afin d'étudier la fréquence d'un mot particulier à différentes époques ou dans différents styles (études synchroniques et diachroniques). Cette méthode originale utilise des équations que l'on peut aisément programmer. L'exemple choisi pour la validation de la méthode est l'étude des variations dans l'emploi des déterminants en grec classique chez Aristophane et Hérodote.

Mots-clés : loi de Poisson, loi binomiale, critère de Kolmogorov, populations parentes, variations linguistiques : synchronie et diachronie.

1. Introduction

Notre méthode a été élaborée dans le cadre d'une étude diachronique de l'article indéfini en grec au cours de laquelle nous avons constaté des variations très nettes dans l'utilisation des morphèmes d'une époque à l'autre (apparition ou disparition de formes, basculements vers un type d'emploi etc.). Pour corroborer nos observations, nous avons mis au point une méthode qui permet de comparer la fréquence d'apparition d'un morphème dans deux textes différents. Après avoir vérifié que la distribution des morphèmes obéit à la loi de Poisson dans chaque texte, nous appliquons un critère qui détermine si les populations sont parentes ou non. Ce test permet d'étudier des variations linguistiques aussi bien dans des corpus de petite taille que dans des corpus de grande taille (dès lors que ceux-ci sont jugés représentatifs pour la démonstration) et est adapté à l'étude de variations de lexèmes dans tous les types de textes. Cette méthode qui a pris naissance dans le cadre d'une recherche en grec peut s'appliquer à toutes les langues.

2. Méthode d'analyse statistique

2.1. *Position du problème*

On se propose d'effectuer une statistique sur la fréquence d'apparition d'un certain mot figurant dans un texte quelconque. Pour réaliser une telle étude, il est nécessaire d'introduire une **mesure** sur les textes, une mesure étant une application sur les nombres réels positifs. La mesure qui semble la plus naturelle et la plus employée repose sur l'ordre des mots comptés à partir du début du texte. Ainsi, la distance de deux mots est la valeur absolue de la différence de leurs rangs. La longueur d'un texte étant évidemment définie par le nombre de mots qu'il contient. À présent, on peut envisager une analyse statistique élémentaire sur la fréquence d'un certain mot X rencontré dans un texte A lequel comprend au total N_a mots. Le mot X est rencontré K_a fois et cela dans les positions notées $x_a(i)$ avec $i = 1, 2 \ldots K_a$.

Le problème que l'on désire résoudre est le suivant : on considère un autre texte appelé B qui contient au total N_b mots, dans lequel on recherche les occurrences du précédent mot X ; celles-ci apparaissent K_b fois et cela dans les positions notées $x_b(j)$ avec $j = 1, 2 \ldots K_b$. À partir de ces données, peut-on conclure à une différence significative ou non des deux populations de X rencontrées dans chacun des textes ? En d'autres termes, a-t-on affaire à la même **population parente** ou à deux populations parentes différentes. L'analyse statistique répond à cette question **sans préjuger des raisons qui peuvent expliquer le fait que les populations sont parentes ou non**.

Notations : dans la suite de ce propos les variables et quantités se rapportant au texte A seront indicées avec a et celles se rapportant au texte B avec l'indice b.

Tel que nous l'avons posé le problème relève de l'analyse statistique des séries d'événements dans laquelle la loi de Poisson joue un rôle fondamental. Soit ξ un nombre d'événements se produisant dans l'intervalle de longueur x. ξ suit une loi de Poisson de moyenne $(\lambda\xi)$, c'est-à-dire :

$$P(\xi = k) = \exp(-\lambda x)(\lambda x)^k / k! \text{ avec } k = 0,1,2...$$

C'est la probabilité que k événements apparaissent dans l'intervalle x. Une propriété importante de la loi de Poisson : les événements qui obéissent à la loi de Poisson sont distribués selon la **loi uniforme** c'est-à-dire que les événements sont équirépartis sur l'axe des abscisses x.

2.2. *hypothèse n° 1 - À l'intérieur d'un texte, les occurrences obéissent à la loi de Poisson*

L'expérience montre que, en règle générale, les occurrences d'un mot obéissent à la loi de Poisson et que cette hypothèse n'a jamais été rejetée lors de l'analyse de plus de *150* cas. Cependant, avant de poursuivre ce calcul, il convient de s'assurer de la validité de cette hypothèse dans chaque cas.

Pour ce faire, point n'est besoin d'estimer le paramètre λ, en effet, il suffit de vérifier que les occurrences sont distribuées uniformément dans le texte. Rappelons que l'on n'obtient jamais de réponse positive à une telle question, **mais on peut savoir si l'hypothèse n'est pas contredite par les données expérimentales**.

2.2.1. *Technique opératoire de la vérification de l'hypothèse N°1*

Il est aisé de construire un **histogramme** des occurrences du texte A. Pour cela il suffit de dénombrer les événements tombant dans chaque intervalle de regroupement. Le nombre I_a

d'intervalles de regroupement (cases) est donné par l'expression :
$I_a = \log_2 (K_a) + 1$, $\log_2$ désignant le logarithme en base *2,* et tous les intervalles de regroupement ont la même taille $h_a = 1/I_a$ (Aïvazian, 1970 ; Ch. Guilpin, 1999). La case $n° j$ contient alors n_j occurrences que l'on est en mesure de dénombrer. Évidemment, $\sum_{j=1}^{I_a} n_j = K_a$, ainsi la probabilité empirique de tomber dans la case $n° j$ s'écrit $p_j = n_j / K_a$. À partir des p_j, il est facile de construire la **fonction de répartition empirique**

$F_n = \sum_{j=1}^{n} p_j$ avec $n = 1, 2 \ldots I_a$, (on s'assure que $F_{I_a} = 1$).

Maintenant, il faut vérifier que l'hypothèse de la répartition uniforme n'est pas contredite par les données expérimentales. Pour cela, **nous allons faire usage du critère de Kolmogorov** pour lequel il convient de déterminer la quantité D_a donnée par l'expression :

$$D_a = \sup |F_n - F_n^*| \text{, où } F_n^* = \frac{n}{I_a} \text{ avec } n = 1, 2, \ldots I_a .$$

À partir de D_a, on calcule la quantité $v_{0a} = \sqrt{K_a} D_a$. v_0 est une valeur possible de la variable aléatoire v laquelle obéit à la loi de Kolmogorov (Ch. Guilpin, 1999). Ainsi, la probabilité pour que v puisse être supérieure à v_0 s'obtient par l'expression :

$$P(v > v_{0a}) = 2 \sum_{k=1}^{\infty} (-1)^{k+1} \exp(-2k^2 v_{0a}^2) .$$

Si la probabilité calculée est inférieure au seuil de signification $\alpha = 0,05$, on rejette l'hypothèse d'une répartition uniforme avec *100* α chances sur *100* de la rejeter à tort. Signalons que le critère de Kolmogorov est sensible à la tendance que le processus étudié s'écarte d'un processus de Poisson (Cox *et al.*, 1969).

Une fois vérifiée l'hypothèse d'une distribution uniforme pour chacun des deux textes considérés appelés *A* et *B*, on cherche à répondre à la question de savoir si le mot *X* est extrait d'une même population parente ou non.

2.2.2. *Comparaison des taux d'occurrence de deux processus de Poisson*

On suppose donc que nous étudions **deux processus de Poisson indépendants** caractérisés par les paramètres λ_a et λ_b que nous cherchons à comparer.

Les deux processus de Poisson sont observés durant les intervalles fixes x_a et x_b avec $x_a = N_a$ et $x_b = N_b$ à l'intérieur desquels on trouve respectivement K_a et K_b occurrences. Ces deux dernières valeurs sont les **valeurs observées des deux variables aléatoires discrètes indépendantes ξ_a et ξ_b**, lesquelles obéissent chacune à une loi de Poisson de moyennes $\mu_a = \lambda_a x_a$ et $\mu_b = \lambda_b x_b$. Il s'ensuit que l'on peut écrire que la probabilité pour que $\xi_a = K_a$ et $\xi_b = K_b$ est donnée par l'expression suivante :

$$(1) \qquad P(\xi_a = K_a , \xi_b = K_b) = \frac{\exp(-\mu_a)\mu_a^{K_a}}{K_a!} \frac{\exp(-\mu_b)\mu_b^{K_b}}{K_b!} .$$

Pour comparer les deux processus, il est commode d'introduire le paramètre ρ ainsi défini :

$$\rho = \mu_a / \mu_b = \lambda_b x_b / \lambda_a x_a \ .$$

Rappelons que x_a et x_b sont des paramètres connus expérimentalement ainsi que K_a et K_b, par conséquent, une inférence sur ρ est équivalente à une inférence sur λ_b / λ_a . À présent, nous devons évaluer la probabilité conditionnelle que $\xi_b = K_b$, sachant que $\xi_a + \xi_b = K_a + K_b$. En définitive, on obtient :

$$P(\xi_b = K_b | \xi_a + \xi_b = K_a + K_b) = \frac{P((\xi_a = K_a).(\xi_b = K_b))}{P(\xi_a + \xi_b = K_a + K_b)} .$$

La probabilité $P(\xi_a + \xi_b = K_a + K_b)$ est donnée par le théorème de la somme de variables poissonniennes indépendantes, à savoir :

$$P(\xi_a + \xi_b = K_a + K_b) = \frac{(\mu_a + \mu_b)^{K_a + K_b}}{(K_a + K_b)!} \exp(-(\mu_a + \mu_b)) .$$

Il en résulte que : $$P(\xi_b = K_b | \xi_a + \xi_b = K_a + K_b) = \frac{\mu_a^{K_a} \mu_b^{K_b}}{K_a! K_b!} \frac{(K_a + K_b)!}{(\mu_a + \mu_b)^{K_a + K_b}} ,$$

puis, en posant $\theta = \frac{\rho}{1+\rho}$, on obtient :

(2) $$P(\xi_b = K_b | \xi_a + \xi_b = K_a + K_b) = C_{K_a + K_b}^{K_a} \theta^{K_b} (1 - \theta)^{K_a} \quad .$$

Cette probabilité est donnée, donc, par la loi binomiale.

2.3. *Hypothèse n° 2 : On se propose de vérifier que les deux lois de Poisson sont les mêmes, c'est-à-dire que* $\lambda_a = \lambda_b$.

Cela signifie que nous devons vérifier que les données expérimentales ne contredisent pas cette seconde hypothèse. Ainsi, si $\lambda_a = \lambda_b$, la loi donnée par l'expression (2) n'est rien d'autre que la loi binomiale de paramètre $\theta = N_b / (N_a + N_b)$.

À partir de (2), il reste à calculer la valeur particulière de ξ_b qui correspond au seuil de signification α que l'on choisit habituellement égal à $0,05$. On désigne par K_{inf} cette valeur limite dont l'usage est le suivant : si $K_b < K_{\inf}$ (pour $\alpha = 0,05$), on rejettera l'hypothèse selon laquelle les deux processus de Poisson relèvent d'une même population parente toujours avec 100α chances sur 100 de rejeter à tort l'hypothèse d'un même processus poissonnien. Dans le cas contraire ($K_b \geq K_{\inf}$), on conserve l'hypothèse qu'il s'agit du même processus poissonnien car elle n'est pas contredite par les données expérimentales ; elle est plausible. Pour déterminer K_{inf}, il suffit d'écrire les deux inéquations suivantes :

(3) $$\sum_{i=0}^{K_{\inf}} C_{K_a + K_b}^{i} \theta^i (1-\theta)^{K_a + K_b - i} \leq \alpha \text{ et } \sum_{i=0}^{K_{\inf}+1} C_{K_a + K_b}^{i} \theta^i (1-\theta)^{K_a + K_b - i} \geq \alpha .$$

Il n'y a aucune difficulté à calculer les valeurs K_{inf} pour la somme $K_a + K_b$ fixée à l'avance.

3. Validation de la méthode : Application à l'étude de deux textes grecs dont les résultats sont connus

3.1. *Introduction*

Cette méthode a été utilisée avec succès sur un très grand nombre de cas empruntés à la littérature grecque. À des fins d'illustrations simples, nous avons retenu des exemples en grec classique (synchronie) dont les résultats sont connus.

Nous étudions des variations dans l'emploi des déterminants du nom en grec ancien. Notre étude est synchronique et porte sur deux genres et deux styles différents au sein du groupe de dialecte ionien-attique. Nous avons choisi deux auteurs de la période classique : Aristophane (vers 445-386 avant J.-C.) qui utilise le dialecte attique du temps, enrichi d'une grande invention verbale, et d'autre part Hérodote (vers 485-425 avant J.-C.) dont le style évolue selon le contexte de la simplicité populaire au ton sentencieux ; son dialecte ionien est parcouru d'emprunts à Homère et à l'attique. Toute proportion gardée, ces deux auteurs sont représentatifs d'une langue qui peut être qualifiée de semi-savante. Pour nos calculs, nous avons utilisé les textes annotés (Habert *et al.*, 1997 et Guilpin, 2003 concernant les problèmes de méthode, codage et ressources concernant le grec) du projet américain Perseus (Tufts University).

3.1.1. *Corpus et données*

The Perseus Digital Library : http://www.perseus.tufts.edu

Aristophanes : *Les Grenouilles* [étude des 1003 premiers vers] - *Aristophanes Comœdiæ*, Hall F.W., Geldart W.M. (éd.), vol. 2, Oxford, Clarendon Press, 1907.

Hérodote : *Histoires* [étude du livre I, du début à LXXV] in Herodotus with an English translation by A.D. Godley, Cambridge, Harvard University Press, 1920.

Le texte *A* écrit par Aristophane comporte 4 873 mots, le texte *B* emprunté à Hérodote contient 11 859 mots. Nous avons $N_a = 4\,873$ et $N_b = 11\,859$.

3.1.2. *Objectifs*

Pour la validation, nous retiendrons la différence de genre entre les deux œuvres : la comédie d'Aristophane suit les conventions de la métrique, tandis que le travail d'historien d'Hérodote est composé en prose de façon à favoriser la mémorisation des textes selon la tradition orale. De ce point de vue, la littérature ionienne offre un support à l'essor intellectuel qui suit et à l'évolution des disciplines (Horrocks, 1997 : 21-3).

Nous nous proposons de vérifier deux phénomènes linguistiques connus en grec ancien a) les variations dans l'emploi du pronom-adjectif indéfini « tis, ti » dont l'apparition est d'autant plus marquée que le texte est savant ou sophistiqué, b) la stabilité des emplois de l'article défini « ho, hê, to » (Biraud, 1991 ; Guilpin, 2002). Nous limitons notre propos à l'étude des GN au nominatif et à l'accusatif. Cette contrainte nous permet par ailleurs de traiter de formes de déterminants communes à l'ionien et à l'attique (Chantraine, 1968).

3.2. *Variations dans l'emploi du pronom-adjectif indéfini « tis, ti »*

Nous partons de l'observation suivante : l'emploi du pronom-adjectif « tis, ti » est d'autant plus fréquent que la langue de l'auteur est savante et son style contraint par le rythme des vers. En revanche, les idiomes plus populaires tendent à effacer la présence de « tis ».

Nous souhaitons donc vérifier l'hypothèse selon laquelle l'emploi du pronom-adjectif « tis » diffère nettement entre l'œuvre versifiée d'Aristophane et l'œuvre en prose d'Hérodote.

On se propose d'analyser la forme neutre du pronom-adjectif indéfini X = « ti ». Précisons que « tis, ti » est traité globalement comme pronom-adjectif dans la mesure où les variations que l'on souhaite mettre en évidence concernent aussi bien les deux formes, notre visée n'étant ni historique, ni épistémologique. Au nominatif et accusatif singuliers. On dénombre alors $K_a = 58$ et $K_b = 25$ occurrences. Il convient de vérifier l'hypothèse de la distribution uniforme pour chacun des deux textes. Il est aisé de calculer que $I_a = 6$ $I_b = 5$ (aux arrondis près). Après avoir réalisé le cumul des fréquences pour obtenir la fonction de répartition, on trouve $D_a = 0,143$ et $D_b = 0,160$ ce qui permet de calculer $v_{0a} = 1,094$ et $v_{0b} = 0,358$. On peut alors obtenir les probabilités correspondantes de pouvoir dépasser ces valeurs : $P_a(v_a > 1,094) = 0,182$ et $P_b(v_b > 0,358) = 0,544$. **Ces résultats indiquent qu'il n'y a aucune raison d'abandonner l'hypothèse d'une répartition uniforme des x_a et x_b.**

Reste à envisager l'hypothèse d'une même loi de Poisson. On calcule alors la valeur. $\theta = N_b/(N_a + N_b) = 0,709$. Il suffit d'utiliser les relations (3) avec la valeur $K_a + K_b = 83$ pour trouver la valeur inférieure de K_b : $K_{\inf} = 52$. La valeur de K_b étant 25, **il nous faut renoncer à l'hypothèse d'une même population parente**.

Il est facile de recommencer la même application avec la forme commune du masculin et du féminin déclinée au nominatif singulier X = « tis » (cas particulier de déterminant « hermaphrodite » en grec). Alors on dénombre $K_a = 20$ et $K_b = 11$, puis on détermine les fonctions de répartition qui permettent d'obtenir $D_a = 0,250$ et $D_b = 0,250$; enfin, on calcule $v_{0a} = 1,118$ et $v_{0b} = 0,829$. À nouveau la probabilité de pouvoir dépasser ces valeurs est dans chaque cas supérieure à $0,05$ car $P_a(v_a > 1,118) = 0,164$ et $P_b(v_b > 0,829) = 0,498$. **On conserve donc l'hypothèse d'une distribution uniforme.**

Pour examiner l'hypothèse d'une seule population parente, il suffit de consulter les tables ou d'effectuer les calculs pour la valeur $K_a + K_b = 31$, on obtient alors : $K_{\inf} = 17$. Ici encore, **on devra renoncer à l'hypothèse d'une unique loi de Poisson** car $K_b = 11$.

Nous avons pu vérifier en terme statistique notre hypothèse : les emplois de « tis » ne sont pas parents. Au cours des siècles, ce phénomène s'accentue jusqu'à la disparition complète du morphème « tis, ti » (gr. byzantin « tinas ») dans la langue démotique. Elle s'amorce dans les dialectes au XV^{e} siècle et prend un tour définitif au XVII^{e} siècle au profit de « enas, mia, ena ». Ce morphème, que l'on trouve sous la forme « heis, mia, hen » jusqu'au X^{e} siècle, désigne en grec ancien le numéral « un » et s'emploie occasionnellement comme pronom indéfini. Dans la langue du Nouveau Testament, il a le statut supplémentaire d'adjectif pronominal à valeur indéfinie. Il ne s'agit pas encore véritablement d'un article. qui acquièrent pendant ce même intervalle de temps toutes les valeurs actuelles de l'article indéfini en grec (Guilpin, 2002). On pourrait vérifier ce phénomène au moyen de notre méthode.

3.3. *Stabilité des emplois de l'article défini*

Originellement, notamment chez Homère, « ho » est un démonstratif, ce qui entraîne dialectalement un emploi comme relatif, puis le mot devient un article, qualifié aujourd'hui de défini, dont les emplois sont bien établis en grec classique (Chantraine, 1968). C'est à partir

de l'examen de ce morphème que Denys le Thrace (170 – 90 avant J.-C.) fonde la catégorie grammaticale de l'article (gr. anc. « arthron », *articulation*, puis lat. « articulus »). Le terme d'*article* lorsqu'il est créé désigne strictement les emplois définis de « ho, hê, to », l'article indéfini n'apparaissant que bien ultérieurement (Guilpin, 2002).

Vérifions que les emplois de « ho, hê, to » sont stables quels que soient la nature du texte étudié et le style de l'auteur.

Intéressons-nous à la forme X = « ton », accusatif masculin singulier de « ho, hê, ton » (le choix de la flexion est arbitraire et le morphème « ton » traité de façon globale). On trouve alors $K_a = 48$ et $K_b = 144$. La détermination des fonctions de répartition fournit les résultats suivants $D_a = 0,055$ et $D_b = 0,0486$ à partir desquels on calcule $v_{0a} = 0,415$ et $v_{0b} = 0,583$. La probabilité de pouvoir dépasser ces valeurs est dans chaque cas supérieure à $0,05$ car $P_a(v_a > 0,415) = 0,99$ et $P_b(v_b > 0,183) = 0,886$. **On conserve donc l'hypothèse d'une distribution uniforme**.

Examinons à présent l'hypothèse d'une seule distribution parente. Il suffit de consulter les tables ou d'effectuer les calculs pour la valeur $K_a + K_b = 192$, on obtient alors pour $K_a : K_{\inf} = 45$. Or K_a vaut 48, **donc il n'y a pas de raison de rejeter l'hypothèse d'une même population parente** (on échange les indices a et b de la relation (2), donc $\theta = N_a / (N_a + N_b) = 0,291$).

Notre hypothèse est donc vérifiée : les emplois de l'article défini sont stables. Cette question va de soi et nous permet d'achever la validation de la méthode.

4. Conclusion

Au cours de la mise au point de cette méthode, nous n'avons jamais pas rencontré de cas où la loi de Poisson devait être rejetée. Toutefois, il est possible qu'un tel cas échoit, ce qui n'interdit pas la poursuite du calcul, mais les résultats devront être traités avec circonspection. On pourra consulter Cox pour l'étude de la comparaison des taux de processus non poissonniens.

Ici, nous avons choisi volontairement des exemples simples et mis en contraste des phénomènes caractéristiques dans la détermination du nom en grec ancien.

Cette méthode pourra être utilisée à des fins plus spécifiques quelle que soit la taille du corpus (dès lors qu'il est jugé représentatif pour la démonstration) et quelle que soit la langue étudiée. De plus, elle permet de traiter aussi bien des données en synchronie qu'en diachronie.

Références

Aïvazian S. (1970). *Étude statistique des dépendances*. MIR.

Biraud M. (1991). *La détermination du nom en grec classique*. Publications de la Faculté de Lettres et Sciences Humaines de Nice.

Chantraine P. (1968-1980). *Dictionnaire étymologique de la langue grecque*. 4 volumes. Klincksieck (reprint 1999).

Cox D.R. et Lewis P.A.W. (1969). *L'analyse statistique des séries d'événements*. Dunod.

Guilpin Ch. (1999). *Manuel de calcul numérique appliqué*. EDP Sciences.

Guilpin P. (2002). Το αόριστο άρθρο στα ελληνικά : Διαχρονική μελέτη (trad. L'article indéfini en grec : étude diachronique). In Clairis Ch. (Ed.), *Recherches en linguistique grecque*, vol. (1), L'Harmattan, *Actes du 5e Colloque International de Linguistique Grecque*.

Guilpin P. (2003). Les textes grecs des origines à nos jours (Ve siècle av. J.-C. – XXIe siècle) – Codage, outils et méthodes de travail. *Lexicometrica* (4).

Habert B., Nazarenko A. et Salem A. (1997). *Les linguistiques de corpus*. Armand Colin/Masson.

Horrocks G. (1997). *Greek : A History of the Language and its Speakers*. Longman.

Jannaris A.N. (1897). *An Historical Greek Grammar, chiefly of the Attic dialect*. Macmillan, 1897 (reprint 1987, Georg Olms : Hidelsheim, Zurich and New York).

Dégrouper les sens : pourquoi, comment ?

Benoît Habert, Gabriel Illouz, Helka Folch

LIR – LIMSI CNRS, BP 133 – 91403 Orsay Cedex – France
{habert,gabrieli,folch}@limsi.fr

Abstract

In order to be able to characterize social points of view in a given domain, a first step consists in spotting words which have several meanings (homonymy, polysemy) or which contexts of use differ widely and in identifying the corresponding shades of meaning. The opposition of use among parts of a corpus is detailed.

Résumé

En lien avec l'objectif global de repérer les points de vue présents dans un domaine, il s'agit de détecter les mots qui ont plusieurs sens ou qui sont employés de manière différente selon les parties d'un corpus et de caractériser leurs emplois. L'application à un corpus partitionné est détaillée.

Mots-clés : sémantique distributionnelle, polysémie, homonymie, repérage de points de vue.

1. Stabilisation du sens et « mondes sociaux »

L'objectif d'un Web sémantique[1], proposé par le W3C, le consortium qui gère le Web, est de modifier la division du travail entre l'homme et la machine dans l'accès aux ressources de la Toile par le sens[2]. Il s'agit d'adjoindre aux contenus informels actuels du Web des connaissances formalisées qui puissent être utilisées par des traitements automatiques et qui permettent ainsi, par exemple, de diminuer le temps passé par les utilisateurs à trier dans les résultats profus voire confus des moteurs de recherches. La proposition du W3C empile les couches (Laublet *et al.*, 2002) : i) expression standardisée des méta-données (Floch et Habert, 2004) ; ii) représentation ontologique convergente ; iii) raisonnement. La première couche entend fournir une syntaxe unifiée pour décrire les ressources et les jugements portés sur elles. La seconde couche a pour objectif un accord sur « ce qui existe », les « étants » manipulés : il faut en effet « parler de la même chose ». C'est le niveau ontologique. La troisième couche englobe les traitements, inférentiels en particulier, possibles à partir du moment où l'on dispose à la fois d'un accord sur l'ontologie à manipuler et d'un méta-langage commun.

Une partie des ressources du Web relève sans doute d'une démarche formalisante poussée, contrôlée par une ontologie partagée, rendant possibles inférences et remodelages et limitant l'intervention humaine. C'est le cas de savoirs techniques et scientifiques stabilisés. Une autre partie des ressources du Web fait coexister des points de vue différents[3] ressortissant à des re-

Ce travail a bénéficié de discussions avec Eric Gaussier (XRCE), André Salem et Serge Fleury (Syled – Université Paris III), Didier Bourigault, Anne Condamines, Cécile Fabre, Josette Rebeyrolle (ERSS – Université de Toulouse-le-Mirail), Elie Naulleau (Semiosys), Claude Henry (LIMSI) et Pierre Zweigenbaum (STIM – AP-HP).

[1] http ://www.w3.org/2001/sw/

[2] Voir aussi http ://www.semanticweb.org/

[3] Le langage de méta-données Topic Maps vise précisément l'articulation de tels points de vue (Floch et Habert, 2002).

présentations semi-formelles où la qualification humaine des informations, l'interprétation, est centrale[4]. La multiplication des forums et des discussions électroniques rend désormais urgent le développement de techniques adaptées à ces fonctionnements sémantiques particuliers[5]. En effet, la désorientation menace aisément face à la multiplication des prises de position dans un domaine donné. La masse même des données fait obstacle à la perception claire des stabilités et des mouvances. Les comptes rendus de réunions publiques et les documents rassemblés par le Centre National du Débat Public[6] sur des thèmes comme la liaison à très haute tension entre l'Espagne et la France[7] sont prototypiques de ces débats citoyens soit directement électroniques soit médiatisés par la mise en ligne.

2. Dégrouper les sens

Un préalable à la mise au jour des points de vue présents dans un ou plusieurs domaine(s) est le dégroupement automatique de sens : le repérage des mots employés simultanément avec des sens divergents au sein du corpus construit pour ce(s) domaine(s). Ces mots recouvent des réalités sémantiques (et sociales) très différentes. Il peut s'agir de simples homonymes : *grêve* 'arrêt du travail'/'plage de gravier', par exemple. Les contextes d'emploi des homonymes sont souvent différenciés. Les différents sens d'un mot polysémique ont plus de chances d'apparaître dans des contextes proches : il y a souvent continuité d'un sens à l'autre, comme pour *guerre*, de *guerre aérienne* à *guerre médiatique*. C'est le cas en particulier des polysémies régulières, de l'engendrement réglé de nouveaux sens : les mots désignant un instrument de musique peuvent ainsi également renvoyer à la personne qui en joue. Un troisième cas de figure — particulièrement pour le discours social — est celui des mots qui incarnent une divergence plus ou moins grande ou un certain « vague », une indétermination du sens (Hanks, 2000). C'est le cas de *service minimum* pour les transports publics.

Au sein de la sémantique « machinale », le dégroupement automatique[8] a cependant, dans l'immédiat, relativement peu retenu l'attention. Les travaux se sont centrés pour l'essentiel sur la désambiguïsation sémantique (*Word Sense Disambiguation*), c'est-à-dire l'attribution en contexte à un mot du sens pertinent en fonction d'un répertoire de sens prédéterminé[9]. En acquisition sémantique, ce sont la mise en évidence de la sous-catégorisation verbale et les regroupements de mots (la recherche de similarités sémantique et leur organisation par classification hiérarchique par exemple) qui ont surtout été explorés (Manning et Schütze, 1999 : ch. 8). Les obstacles sont partiellement techniques : dégrouper les sens, c'est trouver le moyen de repérer les cas où « un mot en cache un voire plusieurs autre(s) », alors que pour les outils de traitements, il s'agit toujours de la même chaîne de caractères. Les obstacles sont également théoriques. La vision « discrétisante » (les sens sont disjoints) et fixiste du sens domine[10]. Paradoxalement, ce sont plutôt des recherches en bibliométrie qui se sont attachées à la comparaison des dénomi-

[4] Imaginer **des** Webs sémantiques, aux fonctionnements distincts, c'est généraliser la conception hétérogène du sens défendue dans Kleiber (1999 : 45-51) : la construction du sens articule plusieurs dimensions qui relèvent de sémantiques partiellement disjointes (instructionnelles, référentielles, etc.).

[5] C'est l'objectif du projet exploratoire RNRT Outiller les alliances (http://www.telecom.gouv.fr/rnrt/index_net.htm) auquel nous avons participé de 2000 à 2003 : fournir des outils pour l'amélioration des débats citoyens organisés par la Fondation pour le Progrès de l'Homme (http://www.fph.ch). C'est aussi celui du logiciel Prospéro (Chateauraynaud, 2003).

[6] http://www.debatpublic.fr/cndp/debat_public_cpdp.html

[7] http://debat-liaison-tht-france-espagne.com

[8] La tâche, dénommée *sense induction* dans Yarowsky (1995), diffère selon qu'on dispose ou non d'une partition préexistante.

[9] Pour un état de l'art, cf. Ide et Véronis (1998) et (Manning et Schütze, 1999 : ch. 6).

[10] *A contrario*, cf. Fuchs et Victorri (1994).

nations par discipline (Losee, 1996) ou aux transferts de dénominations d'un champ à un autre (Losee, 1995) : le mot change de sens en passant ainsi d'un domaine à un autre. Certains travaux en terminologie automatisée ont été consacrés à des phénomènes similaires (Ibekwe-Sanjuan, 1998) : l'évolution d'un groupes de termes au fil du temps par exemple. Le routage d'information, c'est-à-dire l'envoi automatique à l'utilisateur des documents correspondant aux centres d'intérêt qu'il a formulés, rencontre le problème inverse : les mots correspondant aux centres d'intérêt sélectionnés ne sont plus forcément les mêmes.

Malgré les expériences que nous avons déjà menées, notre contribution reste programmatique. Elle vise à présenter de manière raisonnée les différentes facettes du dégroupement de sens : identifier les mots mouvants selon une partition (section 4.1.) ou hors partition (section 4.2.) ; caractériser les directions qui organisent les emplois d'un mot identifié comme mouvant selon une partition (section 4.3.) ou hors partition (section 4.4.). La section 3, liminaire, a pour objectif un ancrage concret et la mise en évidence, par l'exemple, des bénéfices attendus de la direction de travail globale. Pour finir (section 5.1.), nous examinons certaines précautions et certains paramètres à prendre en compte dans la perspective choisie. La section 5.2. conclut la contribution par la question, cruciale mais difficile, des démarches d'évaluation possibles pour les techniques de dégroupement de sens.

3. Mots « mouvants » dans un corpus de textes syndicaux

Les écarts de sens peuvent se manifester par une variation des contextes où figure un mot d'une partie à l'autre d'un corpus. Dans Habert *et al.* (1999)[11], c'est par le biais des fluctuations ou au contraire des stabilités de ces associations <*mot*, *contexte*> que nous essayons de progresser vers le repérage automatique des mots qui font consensus relatif et de ceux qui au contraire témoignent de divergences. C'est donc une *sémantique distributionnelle*. Les associations retenues sont les rapports de dépendance élémentaire entre un « gouverneur » et les mots « pleins » qu'il régit (modifieurs ou arguments) au sein de constituants syntaxiques fournis par des analyseurs syntaxiques robustes.

Le corpus utilisé est celui des résolutions générales (RG) des congrès de la CFTC de 1945 à 1964, de ceux de la CFDT et de la CFTC « maintenue » de 1964 à 1992. La partition utilisée pour contraster les distributions n'est pas basée sur les 3 émetteurs repérables : CFTC originelle, CFDT et CFTC maintenue. Elle repose sur l'interprétation des résultats d'une analyse factorielle des correspondances (Habert et Tournier, 1987) : pour la CFDT, ce sont les événements de 1968 plus que la scission qui entraînent un important changement lexical, suivi d'un autre tournant lexical en 1979. Les parties retenues (correspondant aux agglomérats des deux premiers axes de l'AFC), de taille proche, sont les suivantes : 1) les RG de la CFTC jusqu'à la scission de 1964 incluse et de la CFDT d'avant mai 1968 [TC45-64_DT65-67] (29 704 occurrences) ; 2) les RG de la CFDT maintenue [TC65-90] (31 673 o.) ; 3) les RG de la CFDT « radicalisée » [DT70-76] (25 596 o.) ; 4) les RG de la CFDT « recentrée » [DT79-92] (34 677 o.).

Au sein d'une partie, on peut déterminer pour chaque mot, ses voisins les plus proches. On compare chaque mot aux autres et on détermine pour chaque paire, le nombre de contextes partagés, le nombre de contextes propres à l'un des mots, le nombre de contextes propres à l'autre. Un indice utilise ces quantités et fournit une distance. Les contextes sont les dépendances élémentaires extraites au sein des groupes nominaux fournis par le logiciel d'acquisition terminologique Lexter (Bourigault, 1994). Nous utilisons l'indice de Jaccard[12]. Les voisinages

[11] La revue *Sémiotiques* ayant pris du retard dans ses livraisons, l'article est daté de 1999 alors qu'il est paru en 2002.

[12] D'ailleurs peu approprié puisqu'il « éloigne » les mots qui rentrent dans des nombres de contextes très

observés peuvent différer fortement d'une partie à l'autre. Nous en fournissons deux exemples, où les 5 plus proches voisins d'un mot sont classés par proximité décroissante avec ce mot au sein de la partie :

Pivot	**TC45-64_DT65-66**	**DT70-76**	**DT79-92**	**TC65-90**
action	*lutte, organisation, représentant, participation, syndicalisme*	*lutte, organisation, stratégie, mouvement, pratique*	*lutte, intervention, mobilisation, négociation, revendication*	*mesure, plan, orientation, programme, objectif*
travailleur	*salarié, pays, peuple, classe ouvrière, organisation*	*classe ouvrière, masse, peuple, classe, forces*	*salarié, masse, ensemble, population, syndicat*	*salarié, personne, entreprise, homme, famille*

Les décalages sont très sensibles pour *travailleur* : le voisinage pour TC65-90 est indéniablement d'inspiration chrétienne (*personne, homme, famille*). Le partage de *classe ouvrière* dans les deux premières parties, son absence dans la troisième sont également significatifs. Les voisins sont plus « neutres » pour DT79-92 par rapport aux parties précédentes. Le partage de *lutte* comme voisin d'*action* par les 3 premières parties, mais l'irruption de *négociation* dans la troisième manifestent également des évolutions sémantiques.

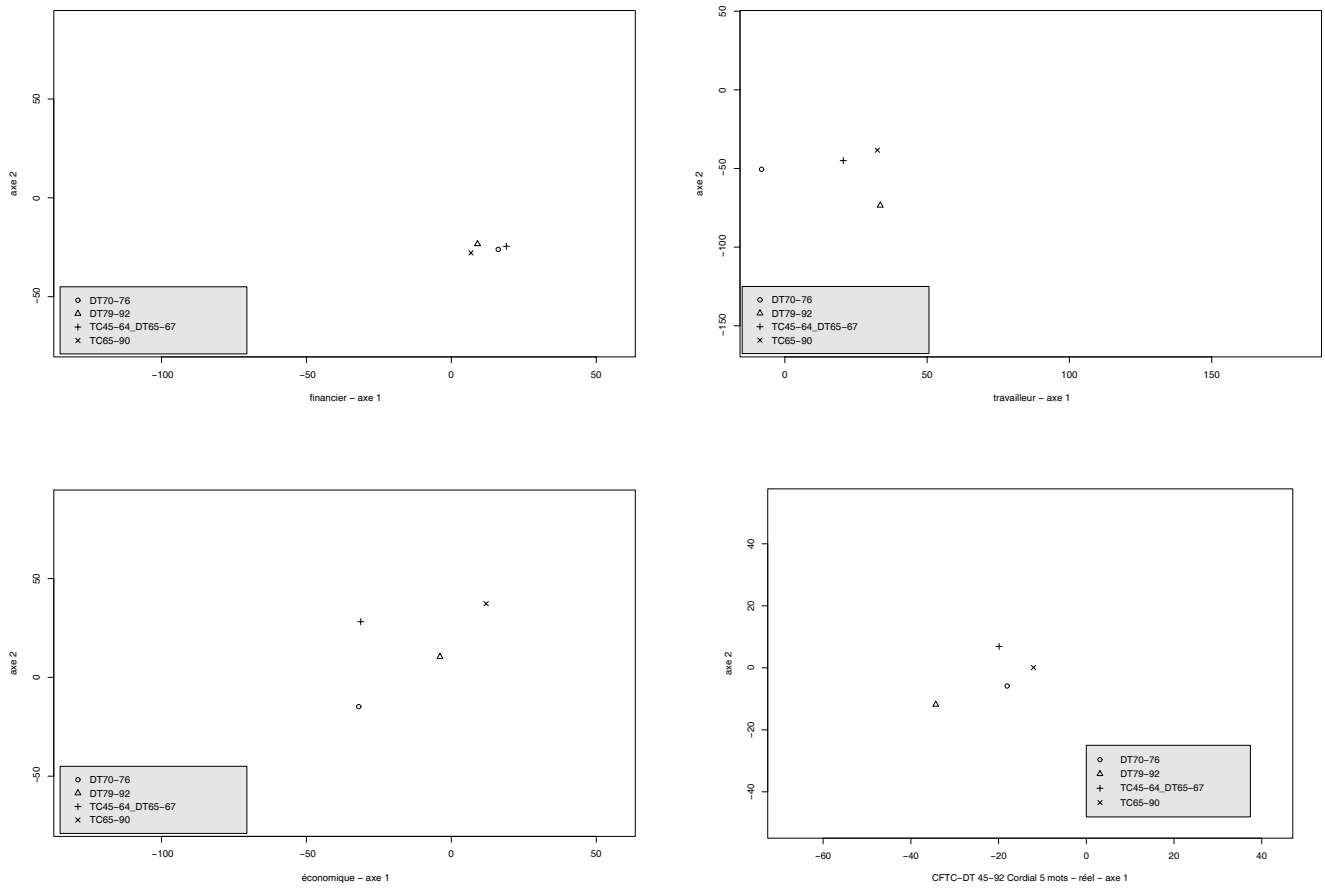
Examiner les voisins d'un mot ne permet pas de détecter les mots dont les contextes changent significativement. Pour avancer dans ce sens, nous retenons, au sein des 100 noms et des 100 adjectifs les plus fréquents dans le corpus, les 26 noms et les 32 adjectifs effectivement partagés par les 4 parties. Dans l'espace à n dimensions[13] de l'union des contextes employés par les k mots partagés considérés, on remplace chaque mot partagé par autant d'hétérographes artificiels que de parties. *Travailleur* est ainsi remplacé par *travailleurDT70-76*, *travailleurTC65-90*… On peut alors examiner les convergences/divergences de contextes selon les parties pour un mot partagé. On utilise une méthode d'analyse multidimensionnelle (Sammon, 1969) pour projeter dans un plan les 4 points correspondant à un même mot. Ces quatre points peuvent être assez ou très rapprochés dans le plan : on peut penser que l'emploi du mot est stable d'une partie à l'autre. Dans le cas contraire, il peut s'agir d'un mot « chahuté ». La figure 1 manifeste une telle opposition entre deux mots pourtant proches sémantiquement « en langue » : *financier* et *économique* (à gauche, de haut en bas). Les emplois d'un mot peuvent faire « bande à part » dans une partie, isolée, alors que les autres sont regroupées. C'est le cas figure 1 de *travailleur* (en haut à droite), où les emplois correspondant à l'époque « radicale » de la CFDT (1970-1967) se distinguent, et de *réel* (en bas à droite) où c'est cette fois la CFDT « recentrée » qui se démarque.

4. Facettes du dégroupement de sens

Les expériences de la section 3 avaient pour objectif de « faire sentir » certaines des dimensions du dégroupement de sens. Nous allons nous appuyer sur elles pour une présentation plus large, qui oppose le repérage des mots mouvants (sections 4.1. et 4.2.) à la caractérisation des (pôles de) sens sous-jacents (sections 4.3. et 4.4.), selon que l'on s'appuie ou non sur une partition du corpus.

différents, même si l'un des mots n'emploie que des contextes utilisés également par l'autre.

[13] 1 400 pour les 26 noms et 763 pour les 24 adjectifs.



4.1. *Identifier les mots mouvants selon une partition*

Des « visualisations » intuitives comme celles de la section 3 permettent un premier repérage des stabilités et mouvances. On peut chercher à leur substituer/associer un indice de dispersion facilitant l'examen et le classement. C'est la démarche de Aussenac-Gilles *et al.* (2003). Les données de l'analyseur syntaxique Syntex (Bourigault et Fabre, 2000) permettent, sur un corpus en ingéniérie des connaissances opposant deux périodes, d'isoler, via des indices appropriés, les termes dont le comportement comme tête/dépendant de groupe syntaxique (nominal ou verbal) a fortement changé d'une période à l'autre. Ces termes sont regroupés manuellement en catégories pour mieux caractériser l'évolution du domaine.

4.2. *Identifier les mots mouvants hors partition*

Dans le cadre d'une représentation vectorielle, le vecteur correspondant à un mot ayant plusieurs sens est l'« assemblage » des vecteurs représentant les sens sous-jacents. Avant de prétendre extraire ces vecteurs sous-jacents (section 4.4.), il faut repérer les vecteurs-assemblages. On peut faire plusieurs hypothèses. Dans une perspective proche de celle de Salton *et al.* (1997), il est possible que les mots à sens multiples contribuent à rapprocher les unités textuelles utilisées (documents, phrases)[14]. Il s'agit alors de détecter les mots qui, lorsqu'on les enlève, laissent place à un « nuage » d'unités textuelles plus dilaté (en recherche d'information, ce sont au contraire les mots qu'on souhaite éliminer). Une autre hypothèse est qu'un mot à sens multiple

[14] En analyse factorielle des correspondances, ce serait un sous-ensemble des formes qui contribuent **le moins** à créer les oppositions majeures.

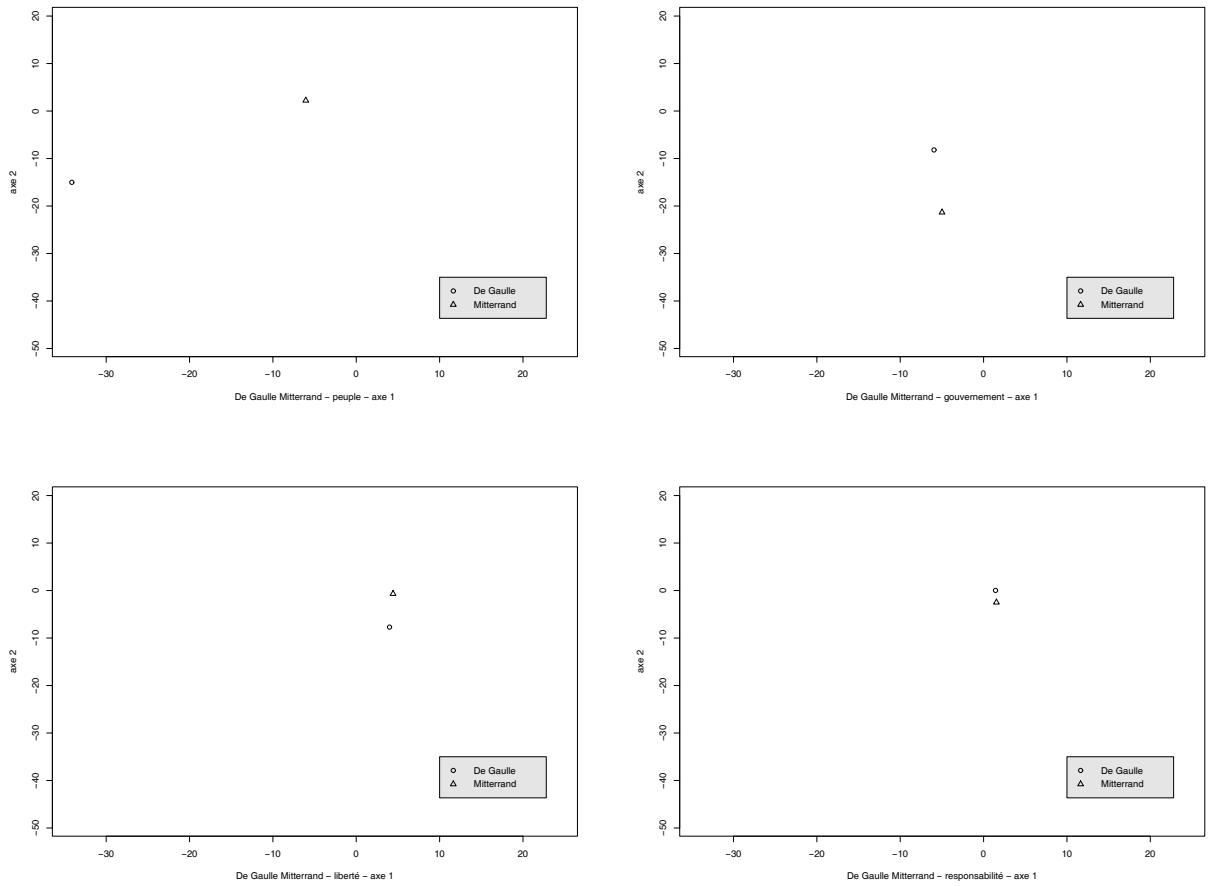
aurait des voisins moins proches entre eux qu'un mot plus univoque. C'est le développement de l'intuition exemplifiée pour *travail* et *action* en section 3. Il s'agit alors de développer des métriques de dispersion du nuage des voisins, en tenant compte de la répartition inégale du nombre de voisins selon le mot examiné.

4.3. *Contraster les emplois selon une partition*

Le dégroupement de sens est dans ce cadre guidé (*supervisé*) par une partition préexistante qui fournit le cadre d'observation et de détection des contrastes distributionnels.

Cette partition peut reposer sur une signalétique externe : datation, émetteur, etc. C'est la démarche suivie par Folch (2002). Dans le cadre du projet *Scriptorium*, consacré au sein de la Direction des Etudes et Recherches d'EDF à l'étude des débats autour de thématiques internes à l'entreprise, un corpus a été constitué autour de la notion de *service public* et de sa rédéfinition avec l'ouverture à l'Europe. L'opposition entre émetteurs (les différents syndicats, la direction) fonde la partition au sein de laquelle sont observées, par les méthodes exposées en section 3, les convergences et divergences d'emploi des mots. Nous avons appliqué également ces méthodes à la mise en regard des interventions radio-télévisées de De Gaulle et de Mitterrand[15]. On observe ainsi sur la figure 2 un grand écart entre De Gaulle et Mitterrand dans l'emploi de *peuple* (en haut à gauche), ce qui ne surprend guère, et une grande proximité dans celui de *gouvernement* (en haut à droite), ce qui n'étonne pas non plus, tandis que le rapprochement sur *liberté* (en bas à gauche) et *responsabilité* (en bas à droite) est plus inattendu. Nous avons également examiné quelques pôles (cf. figure 3) dans les discours de Mitterrand en opposant la cohabitation avec ce qui précède (en reprenant le partage proposé par D. Labbé). La proximité sur *Europe* (en haut à gauche) s'oppose à l'éloignement pour *France* (en haut à droite). Il en va de même pour *gouvernement* (en bas à gauche) et *ministre* (en bas à droite). On retrouve en filigrane l'opposition entre le premier ministre et le président de la France, de la cohabitation, alors que ni l'Europe ni le gouvernement ne sont présentés de manière sensiblement différente.

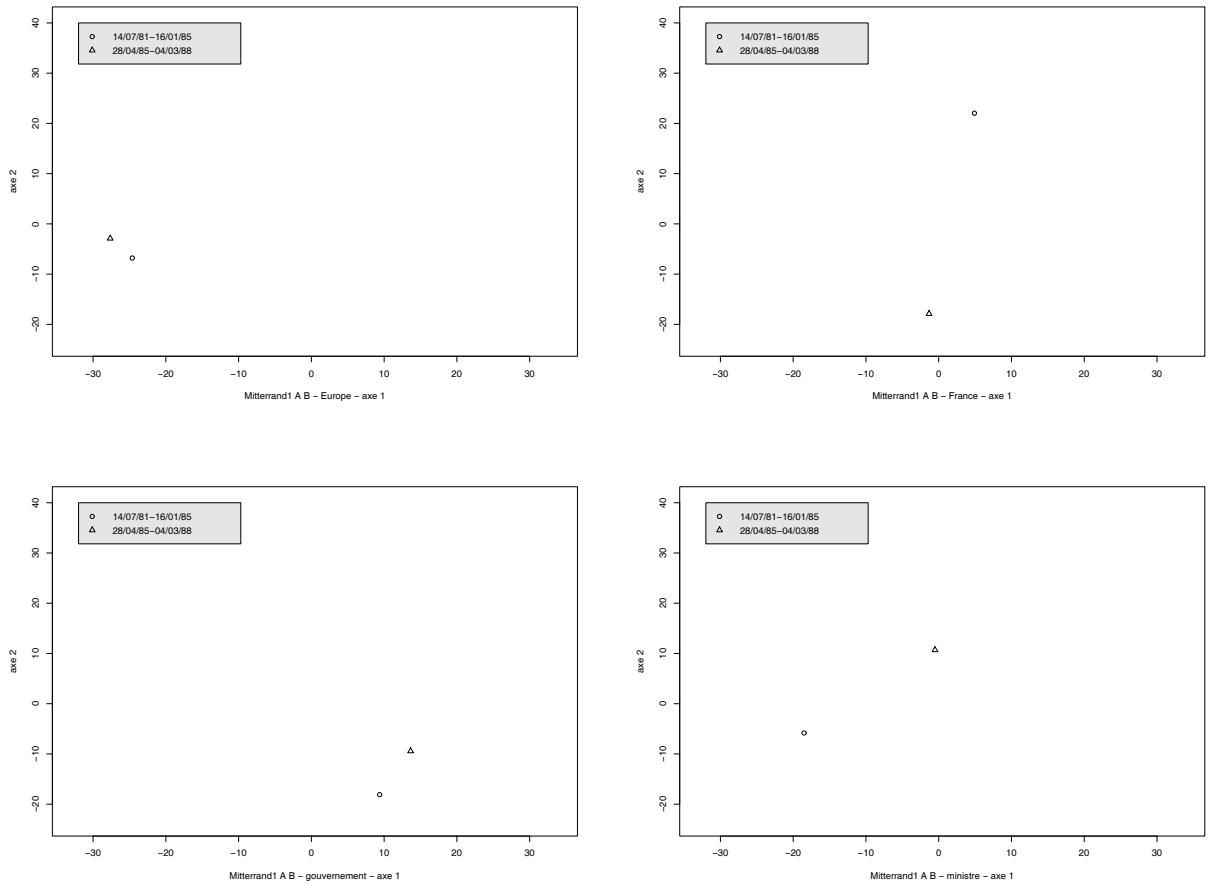
[15] Le corpus nous a été fourni par D. Labbé, que nous remercions.



La partition peut résulter d'outils de partitionnement (*clustering*) qui regroupent les unités textuelles choisies sur la base des traits jugés pertinents (formes brutes, formes racinisées – *stemming*, lemmes, lemmes moins mots-outils, etc.). Le corpus peut imposer cette démarche. Les archives d'un forum électronique peuvent ainsi permettre une répartition par émetteur (auteur) qui s'avèrera en réalité inutilisable si l'« émiettement » en une multitude d'auteurs rend imperceptibles les fluctuations des associations. C'est une démarche de regroupement préalable de cet ordre qui a été suivie dans l'expérience rapportée en section 3. Le partitionnement par interprétation des résultats d'une analyse factorielle préalable permettait en particulier de « lisser » les décalages de taille importants entre les documents utilisés (de 357 à 14 100 o.). Il aboutissait également à un nombre « raisonnable » de parties, où les contrastes et ressemblances étaient plus aisément saisissables. Dans Sébillot (2002), les 9 500 paragraphes d'un corpus de 200 articles du Monde diplomatique (des années 1987-97) sont représentés par les 165 noms les plus fréquents. Une classification hiérarchique de ces 165 noms débouche sur 80 classes thématiques, réduites aux 27 jugées pertinentes (et « nommées ») par l'accord de 4 évaluateurs sur les 5 mis à contribution. Si deux mots d'un thème sont présents dans un paragraphe, le paragraphe est placé dans ce sous-corpus thématique. Un paragraphe peut apparaître dans plusieurs thèmes[16]. Au sein de chaque thème, les noms les plus fréquents sont classés. Les traits sont les noms, verbes et adjectifs dans une fenêtre de plus ou moins 5 mots. Il est possible alors d'examiner pour un mot donné ses voisins dans deux thèmes, les partages et les divergences. Lorsqu'on considère les thèmes Négociations et Territoire, les voisins *état, parti, économique, nouveau, place, arrivée, politique* sont partagés, tandis que *accession, an, armée, concentra-*

[16] Les 27 classes thématiques ne constituent donc pas une partition de l'ensemble des paragraphes.

tion, pays, coalition, contrôle, gouvernement, partage, achat, central, public sont propres au premier thème et *local, soviétique, année, exécutif, prise, public, président, central* au second. Une visualisation des rapprochements/divergences serait probablement plus confuse avec au moins théoriquement jusqu'à 27 points pour un mot donné.



Au total, recourir à une partition « donne à voir » des oppositions. En revanche, cela tord éventuellement les contrastes, puisqu'on ne peut plus percevoir les décalages internes à une partie ni d'autres oppositions et rapprochements qui naîtraient d'autres partitions possibles (par exemple pour le corpus syndical de la section 3 entre CFTC maintenue, CFDT et CFTC jusqu'en 1964). C'est pourquoi il est sans doute préférable dans tous les cas de « rebattre les cartes », en testant plusieurs partitions, mais aussi en rassemblant hors partition (section 4.4.) les contextes d'un mot détecté comme « mouvant » et en les soumettant, de manière éventuellement couplée (Lebart *et al.*, 1997 : 185-206), aux techniques de classification et d'analyse de correspondances.

4.4. *Contraster les emplois hors partition*

Rapp (2003) part de l'hypothèse que ce qui est appelé un mot ambigu est l'« assemblage » des vecteurs représentant les sens sous-jacents : le vecteur normalisé correspondant à *bank* devrait être plus similaire à celui résultant de la somme des vecteurs normalisés pour *money* et *river* qu'à la somme des vecteurs normalisés de toute autre paire de mots. Rapp cherche les vecteurs sous-jacents aux 12 mots dont la désambiguïsation est examinée dans Yarowsky (1995) (*axes : grid/tools ; bass : fish/music ; crane : bird/machine ; drug : medicine/narcotic ; duty : tax/obligation ; motion : legal/physical ; palm : tree/hand ; plant : living/factory : poach : steal/boil ; sake : benefit/drink : space : volume/outer ; tank : vehicle/container*). Il se can-

tonne à deux sens possibles seulement par mot ambigu[17]. Une fenêtre de ± 1 mot plein, sur les 100 millions des mots du *British National Corpus* (BNC), fournit les contextes sous-jacents aux vecteurs. L'algorithme cherche les 10 mots les plus fortement associés à un mot ambigu (les contextes « majeurs »). Il produit les 90 paires possibles issues de ces 10 mots. Il calcule enfin la similarité entre ces paires et le mot ambigu. La plus ou moins grande similarité entre les vecteurs est mesurée par la distance de Manhattan. Pour *bank*, les 10 premières associations, par similarité décroissante, sont *account/river, accounts/river, accounts/manager, account/manager, account/accounts, loans/river, accounts/holiday, account/loans, account/holiday, central/account*. Malgré l'absence de lemmatisation, qui produit des néo-doublons (*account/river ; accounts/river*), les paires renvoient pour l'essentiel aux deux sens de *bank*. Pour les 2 premières paires correspondant aux 12 mots choisis, 9 paires sur 24 (37.5%) correspondent aux sens distingués par Yarowsky. Il est également possible d'appliquer des méthodes de classification aux contextes associés aux mots ambigus, avec un risque : *les grands regroupements (top-level cluster partitions) basés uniquement sur l'information distributionnelle n'entrent pas forcément en correspondance avec les acceptions généralement reconnues* (Yarowsky, 1995 : 195).

5. Paramétrages, précautions et évalution

5.1. *Paramètres et précautions*

Dans le cadre d'une sémantique distributionnelle, un obstacle partagé par les différentes tâches (acquisition, désambiguïsation) est l'émiettement et le déséquilibre des distributions des sens d'un mot. A titre d'exemple, les 1 345 phrases contenant *vendre* ou un de ses dérivés extraites du corpus PAROLE de 14 millions de mots de numéros extraits aléatoirement des années 1987, 1989, 1991, 1993 et 1995 du journal *Le Monde* ne contenaient aucun exemple du sens 'trahir', pourtant présent dans tous les dictionnaires. Le dégroupement de sens bute sur le même obstacle. Dans Rapp (2003), qui s'appuie pourtant sur les 100 millions de mots du BNC, les deux premières paires pour *bass* (*fish/music*) sont *guitar/treble* (soprano, clef de sol) et *guitar/string* : ce échec est attribué par Rapp à la mauvaise représentation du sens *fish* dans le corpus. Il en va de même pour *tâche* dans Aussenac-Gilles *et al.* (2003) dont les deux sens ('structure de représentation'/'tâche prescrite') sont très inégalement représentés. Cela doit conduire à veiller non seulement à une taille minimale des corpus utilisés mais à leur diversité thématique. Il est par ailleurs possible (section 4.3.) qu'un nombre limité de parties facilite les mises en évidence de sens divergents.

Par ailleurs, la répartition des traits permettant de classer les mots est souvent très éparpillée : les matrices résultantes sont fortement creuses. Une des approches possibles est le calcul de similarités de second ordre (Grefenstette, 1994a). Dans Sebillot (2002), les mots pleins qui servent à classer les noms sont ainsi remplacés dans un deuxième temps par des regroupements de ces mots pleins. On passe d'une matrice 383x8 000 à une matrice 383x544 : l'espace des traits est divisé par 15.

Dans l'esprit de Grefenstette (1996) et de Curran et Moens (2002), on peut chercher la définition la plus opératoire des traits utilisés, les contextes, en faisant varier leur taille (empan limité de k mots / phrases, paragraphes) et leur nature (formes graphiques, lemmes, positions syntaxiques) et en tenant compte de la plus ou moins grande adéquation des textes traités aux outils de segmentation, d'étiquetage ou de structuration disponibles. Les deux contraintes à concilier

[17] Rapp pense que l'algorithme se généralise à 3… k sens. On peut en douter. En premier lieu, augmenter les sens multiplie les tuples à examiner et atténue les écarts entre la somme des vecteurs de ces tuples. En second lieu, savoir le nombre de sens effectivement présents reste un problème à part entière, à supposer même qu'un tel but soit accessible.

sont d'une part de disposer de suffisamment de contextes pour rapprocher/différencier les mots de manière fiable (ce qui privilégie les contextes larges et les formes graphiques) et d'autre part de bénéficier de contextes plus précis et aisément interprétables (ce qui avantage les collocations restreintes et les dépendances syntaxiques). On peut d'ailleurs découpler les deux problèmes, en utilisant des contextes larges et « grossiers » pour repérer les mots mouvants/stables sur le plan sémantique et en utilisant des contextes restreints et plus parlants pour aider à l'interprétation.

Enfin, la constitution des unités textuelles au sein desquelles opérer les dégroupements peut se révéler délicate. Dans bien des cas, les unités « naturelles » que sont le document ou le paragraphe conviennent. Il peut s'avérer nécessaire néanmoins de disposer d'un grain plus fin. Un compte rendu de débat suppose d'isoler les tours de parole et de les rattacher à des émetteurs. Le mécanisme des reprises et réponses dans un mail inséré dans un fil de discussion de forum implique éventuellement de rattacher des fragments d'un mail donné à des émetteurs distincts, pour éviter les fausses proximités liées aux propos qu'un acteur rapporte mais qu'il ne prend pas à son compte. Le nombre parfois élevé des intervenants dans un forum ou dans un débat peut demander en aval de cet éclatement de regrouper les émetteurs en catégories, pour que les variations de sens soient tout simplement perceptibles.

5.2. *Évaluer ?*

La désambiguïsation sémantique a permis le développement de méthodes d'évaluation, via les campagnes SensEval (Kilgariff et Palmer, 2000). Un répertoire de sens prédéterminé, basé sur un dictionnaire existant, sert à étiqueter manuellement — dans le corpus d'entraînement et dans le corpus sur lequel seront jugés les systèmes en compétition — l'ensemble, relativement restreint, de mots à désambiguïser. En acquisition de catégories sémantiques, le rattachement de deux mots à une classe (*cluster*) est considéré comme juste s'ils figurent dans une même catégorie de thesaurus (Grefenstette, 1994), thesaurus qui peuvent être éventuellement combinés (Curran et Moens, 2002).

L'emploi de dictionnaires ou de thesaurus existants pour le dégroupement de sens peut permettre, pour des mots « de langue générale », d'examiner la corrélation entre le nombre de sens distingués dans ces ouvrages de référence et le « degré de mouvance » fourni par une méthode développée pour le dégroupement[18]. Ces références[19] ne conviennent plus forcément pour les divergences de point de vue en veille sociale. Une évaluation a posteriori des listes de mots univoques / « mouvants » fournies par les algorithmes est peu fiable : les facteurs conditionnant la satisfaction ou l'insatisfaction observées sont peu contrôlables. La présence d'experts du domaine peut permettre d'envisager une comparaison des listes obtenues avec les jugements formulés a priori par ces experts sur les mots les plus fréquents du corpus. On risque néanmoins de se heurter au paradoxe mis en évidence par les expériences rapportées par Véronis (2004) : la tâche de répérage de mots polysémiques semble facile aux annotateurs mais elle débouche sur de faibles taux d'accord.

Comme Yarowsky, on peut aussi engendrer des « pseudo-mots » : fusionner en une étiquette arbitraire deux mots dont on sait qu'ils étiquètent le sens de deux homonymes ou d'un mot polysémique. C'est inverser la démarche de Rapp : engendrer *bank-river*, *grid-tools*, etc. On

[18] Bien que le traitement de la polysémie et de l'homonymie varie sensiblement d'un dictionnaire à l'autre, la ligne de partage devra être conservée. La polysémie est sans doute plus difficile à détecter. On notera d'ailleurs que les 12 mots ambigus traités dans Rapp (2003) correspondent à des homonymes. Les résultats se dégraderaient probablement avec des mots polysémiques.

[19] Malheureusement non aisément accessibles pour la recherche sur le français (à la différence de WordNet et de multiples thesaurus en ligne pour l'anglais).

peut alors examiner si ces pseudo-mots sont effectivement repérés comme « mouvants ».

5.3. *Interpréter*

Les distinctions sémantiques en dégroupement de sens sont labiles et sujettes à caution. Dans le même esprit que Aussenac-Gilles *et al.* (2003), il nous paraît crucial de ne pas céder aux mirages d'indices opaques et de revenir systématiquement aux contextes, en s'appuyant sur des architectures de gestion de corpus et de traitement adaptées (Folch, 2002).

Références

Aussenac-Gilles N., Bourigault D. et Teulier R. (2003). Analyse comparative de corpus : cas de l'ingéniérie des connaissances. In *14èmes journées francophones d'ingénierie des connaissances (IC 2003)* : 67-84.

Bourigault D. (1994). *LEXTER un Logiciel d'EXtraction de TERminologie. Application à l'extraction des connaissances à partir de textes*. Thèse en mathématiques, informatique appliquée aux sciences de l'homme. Paris, École des Hautes Études en Sciences Sociales.

Bourigault D. et Fabre C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de grammaire*, vol. (25) : 131-151.

Chateauraynaud F. (2003). *Prospéro : une technologie littéraire pour les sciences humaines*. CNRS Éditions.

Curran J.R. et Moens M. (2002). Scaling context space. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)* : 231-238.

Folch H. (2002). *Articuler les classifications sémantiques induites d'un domaine*. Doctorat en informatique. Université Paris XIII.

Folch H. et Habert B. (2002). Articulating conceptual spaces using the Topic Maps standard. In Wood L. (Éd.), *Proceedings XML 2002*.

Folch H. et Habert B. (2004). Langages de méta-données pour le Web sémantique : RDF et Topic Maps. In Ihadjadène M. (Ed.), *Outils et méthodes en recherche d'information*. Hermès. À paraître.

Fuchs C. et Victorri B. (Eds) (1994). Continuity in linguistic semantics. *Lingvisticae Investigationes Supplementa*, vol. (19). John Benjamins.

Grefenstette G. (1994a). Corpus-derived first, second and third order affinities. In *Proceedings of EURALEX*.

Grefenstette G. (1994b). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publisher.

Grefenstette G. (1996). Evaluation techniques for automatic semantic extraction : Comparing syntactic and window based approaches. In Boguraev B. and Pustejovsky J. (Eds), *Corpus Processing for Lexical Acquisition. Language, Speech and Communication* : 205-216.

Habert B., Folch H. et Illouz G. (1999). Sortir des sens uniques : repérer les mots « mouvants » dans le domaine social. *Sémiotiques*, vol. (17). *Dépasser les sens iniques dans l'accès automatisé aux textes*, Habert B. (resp.) : 121-151.

Habert B. et Tournier M. (1987). La tradition chrétienne du syndicalisme français aux prises avec le temps. Évolution comparée des résolutions générales CFTC, CFDT et CFTC-maintenue (1945-1985). *MOTS. Presses de la Fondation Nationale des Sciences Politiques*, vol. (14) : 21-46.

Hanks P. (2000). Do word meanings exist ? *Computers and the Humanities*, vol. (34/1-2) : 205-215.

Ibekwe-SanJuan F. (1998). Terminological variation, a means of identifying research topics from texts. In *Proceedings of ACL-COLING'98* : 654-661.

Ide N. et Véronis J. (1998). Introduction to the special issue on word sense disambiguation : the state of the art. *Computational Linguistics*, vol. (24/1) : 1-40.

Kilgariff A. et Palmer M. (Eds) (2000). Senseval : Evaluating Word Sense Disambiguation Programs. *Computers and the Humanities*, vol. (34). Kluwer.

Kleiber G. (1999). *Problèmes de sémantique : la polysémie en question. Sens et structures*. Presses Universitaires du Septentrion.

Laublet P., Reynaud C. et Charlet J. (2002). Sur quelques aspects du Web sémantique. In *Actes du GDR I3*.

Lebart L., Morineau A. et Piron M. (1997). *Statistique exploratoire multidimensionnelle*. Dunod.

Losee R.M. (1995). The development and migration of concepts from donor to borrower disciplines : Sublanguage term use in hard and soft sciences. In *Proceedings of 5th International Conference on Scientometrics and Informetrics* : 265-274.

Losee R.M. (1996). Text windows and phrases differing by discipline, location in document, and syntactic structure. *Information Processing and Management*, vol. (32/6) : 747-767.

Manning C.D. et Schütze H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.

Rapp R. (2003). Discovering the meanings of an ambiguous word by searching for sense descriptors with complementary context patterns. *Terminologie et Intelligence Artificielle* : 145-155.

Salton G., Wong A. et Wang C.S. (1997). A vector space model for automatic indexing. In Sparck Jones K. et Willett P. (Eds), *Readings in Information Retrieval* : 273-289. Article publié en 1975.

Sammon J. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computing*, vol. (18) : 401-409.

Sébillot P. (2002). *Apprentissage sur corpus de relations lexicales sémantiques. La linguistique et l'apprentissage au service d'applications du traitement automatique des langues*. Habilitation à diriger des recherches. Université de Rennes I. (IRISA Documents d'habilitation 41).

Véronis J. (2004). Quels dictionnaires pour l'étiquetage sémantique ? In Fuchs C. et Habert B. (Resp.), *Le français moderne. Traitement automatique des langues et linguistique*.

Yarowsky D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting*. Association for Computational Linguistics : 189-196.

Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex

Serge Heiden

ICAR UMR5191 – ENS-LSH – 69342 Lyon BP7000 Cedex 07 – France
slh@ens-lsh.fr

Abstract

There are a lot of different statistical tests that can help us to judge if two different words do co-occur or not in a corpus. Instead of trying to give a significance level to a particular test on linguistic data, we try to use it to explore the co-occurrence space through the use of an hypertext interface. We use a co-occurrence model conceived in our laboratory. We have implemented it in our tool called *Weblex*. Through the example of a corpus of speeches from different speakers of French "Assemblée constituante", we illustrate the to and fro movement available. This is done through different levels of synthesis with the ply between concordances, lexicograms and recursive lexicograms. Finally we show that it is the dynamic nature of skimming that permits to build the interpretation. Future work will concern better semi-automatic parameter tuning and on the fly annotation of information.

Résumé

Il existe de nombreux indices statistiques permettant de juger de l'attirance entre deux mots dans un corpus de textes. Plutôt que d'accorder une valeur de vérité à un indice particulier dans le domaine linguistique, nous cherchons à l'utiliser au mieux afin d'explorer l'ensemble de l'espace de cooccurrence au moyen d'une interface hypertextuelle. Nous exploitons un modèle de calcul direct de cooccurrence propre à notre laboratoire. Nous avons implémenté ce modèle dans notre outil *Weblex*. Sur la base d'un corpus exemple de discours d'orateurs de l'Assemblée constituante, nous illustrons le va-et-vient offert dans l'espace de cooccurrence. Ceci est réalisé à travers différents niveaux de synthèse possibles dans un parcours successif à travers des concordances, des lexicogrammes et enfin des lexicogrammes récursifs. Finalement nous montrons que c'est la dynamique du parcours qui permet de construire l'interprétation. Les développements futurs concernent de nouveaux asservissements de réglage de paramètres ainsi que des possibilités d'annotation d'informations en cours de parcours.

Mots-clés : cooccurrence, concordance, lexicogramme, lexicogramme récursif, hypertexte, Weblex.

1. Introduction

Il existe une vaste littérature sur les différents modèles de cooccurrences disponibles et sur leurs multiples applications : en lexicographie, analyse de discours, génération de texte, analyse syntaxique, etc. (Manning *et al.*, 2002 ; Oakes, 1998). L'indice *t* permet, par exemple, d'évaluer si la rencontre entre deux mots est due au hasard ou non. Il permet encore d'obtenir les cooccurrents les plus discriminants entre deux mots. L'information mutuelle permet de comparer la probabilité que deux mots cooccurrent (se rencontrent) avec celle de l'apparition indépendante de chaque mot. La log-vraisemblance, l'écart réduit Z, le test du Khi^2, l'Entropie, etc. sont tous susceptibles à leur manière de fournir des informations utilisables dans ce domaine. Dans cet article nous présentons l'usage que nous faisons d'un test de calcul direct de probabilité de cooccurrence conçu dans notre laboratoire (Lafon, 1984), dans la même lignée que le test de Fisher. Nous ne cherchons pas à établir si tel test ou tel autre est particulièrement « significatif » sur des données du domaine linguistique (Manning *et al.*, 2002 :

166), ou à leur accorder une interprétation linguistique directe comme certains cherchent à le faire (Krenn, 2001), voire pire à leur donner une valeur de vérité dans le domaine de l'attribution d'auteur (Labbé, 2001). Plus modestement, nous cherchons à exploiter un indicateur nous permettant de repérer des attirances entre mots de toutes natures : lexicales quand la distance moyenne et l'écart-type entre mots sont faibles, syntaxiques pour les distances plus grandes, thématiques au delà, etc. Comme la quantité d'informations sur les cooccurrences dans les grands corpus est potentiellement énorme, nous allons construire différentes vues sur cet espace de cooccurrence au moyen d'un hypertexte. De nombreux paramètres vont par ailleurs nous permettre de régler le type d'attirance mis en évidence : tris supplémentaires sur la distance moyenne, choix d'élagage systématique dans le vocabulaire considéré, choix du contexte de cooccurrence, des lexies prises en compte (lexème et/ou locutions), etc. L'objectif de cet article est donc de présenter les enjeux de l'exploitation d'un indice de cooccurrence particulier dans le cadre d'une interface informatique de type hypertexte.

Après avoir présenté le modèle utilisé dans la section 2, nous présenterons à la section suivante l'usage hypertextuel en couches de synthèses que nous faisons de la probabilité de cooccurrence dans l'outil *Weblex*. Enfin, en conclusion nous ferons la synthèse des enjeux d'interfaçage de ce parcours et nous présenterons quelques développements futurs.

2. Un modèle de cooccurrences reposant sur un calcul direct

Le modèle présenté dans cet article est implémenté dans l'outil *Weblex* développé au laboratoire ICAR de l'Ecole Normale Supérieure de Lettres et sciences humaines (Heiden, 2002). La mise en œuvre est réalisée en deux phases : tout d'abord une phase de préparation du corpus permet d'expliciter au fil du corpus un certain nombre d'informations qui seront exploitées dans la phase suivante (les lexies et les contextes) ; ensuite une phase d'analyse met en œuvre le modèle sur la base de la préparation. Pour des raisons de place, nous n'exposerons que le modèle lui-même suivi de la deuxième phase de sa mise en œuvre dans l'outil.

Le modèle de cooccurrences de Weblex

Nous nous situons dans le cadre où deux phénomènes linguistiques se rencontrent dans le même contexte sans être obligatoirement juxtaposés[1]. Le phénomène considéré dans cet exposé est le lexème, c'est à dire un mot simple, alors que n'importe quel phénomène, en fonction de la phase de préparation, peut être supportée par le modèle (locution, propriété d'occurrences : comme les parties du discours, lemme, etc.). Nous ne considérerons par ailleurs que la succession (orientée donc) de ces phénomènes que nous désignerons par « couple de cooccurrents », alors que les paires de mots (non orientées) font l'objet d'un autre modèle. Enfin le contexte de rencontre sera la phrase orthographique (essentiellement délimitée par la ponctuation forte : ., !, ?, …), alors que d'autres contextes sont envisageables comme le syntagme, la proposition, le paragraphe, etc.

Le modèle utilisé est celui de Lafon (1984). Étant donnés : les fréquences totales f_A et f_B de deux mots A et B du texte et V son vocabulaire, le calcul direct de la probabilité que ces deux mots se rencontrent exactement r fois dans les P phrases du texte est donnée par (la flèche dénotant la succession dans une même phrase) :

$$P(card\{A \in Vet B \in V / A \rightarrow B\} = r) = \frac{C_r^{fA} \times C_{fB-r}^{P+fB}}{C_{fB}^{fA+fB+P}}$$

[1] Ceci justifie l'appellation de cooccurrences en opposition aux collocations classiquement juxtaposées.

L'indice utilisé dans *Weblex* est finalement celui de la probabilité que ces deux mots se rencontrent le nombre de fois qu'on le constate effectivement dans le texte et plus encore, à concurrence de la fréquence minimum des deux mots. Cette probabilité est obtenue en sommant les valeurs de la probabilité discrète sur l'ensemble de la queue de la distribution.

L'implémentation efficace en langage C de cet indice repose sur une double série de fonctions rémanentes[2], ainsi que sur un agencement précis de l'ordre des opérations arithmétiques afin d'éviter tout phénomène d'underflow.

3. Mise en œuvre hypertextuelle du modèle

Dans cet article nous illustrerons un parcours descriptif basé sur un corpus historique fermé constitué de discours des orateurs de l'Assemblée constituante. N'étant pas historien nous-même, nous ne chercherons pas à confirmer ou à infirmer une hypothèse de travail mais plutôt à décrire l'analyse en cooccurrences de divers fonctionnements discursifs. Donc plutôt que l'aspect probatoire de la méthode ce sera son aspect heuristique — de découverte par la lecture transversale du corpus — que nous illustrerons.

Comme nous utiliserons comme indice quantitatif fondamental la fréquence de chaque forme du vocabulaire de chaque orateur dans nos analyses, présentons d'abord les limites quantitatives globales du corpus à travers le tableau des dimensions lexicométriques pour chaque locuteur (voir figure 1).

Corpus	**Occurrences**	**Formes**	**Phrases**
Mirabeau	96404	8865	4120
Sieyès	13875	2233	677
Total	110279		4797

Figure 1.Dimensions lexicométriques du corpus des orateurs.

La colonne *corpus* indique l'orateur concerné, *occurrences* le nombre de mots, *formes* la taille du vocabulaire de l'orateur et *phrases* le nombre de phrases orthographiques.

3.1. *Réalisations de la base « constitution »*

Prenons comme objet d'étude le champ sémantique de *constitution* tel que la surface graphique du discours de chaque locuteur nous permet de l'appréhender. Le choix de ce champ est fortuit, on procède selon la même méthode pour n'importe quel champ notionnel.

Indépendamment des différentes formes que pourra prendre la base « constitution » dans ce corpus, l'interprétation de chacune de ses réalisations ne peut, au final, s'effectuer qu'en contexte, c'est à dire en situant chaque mot dans le discours, par exemple la phrase où il apparaît. Le calcul classique des *Contextes* d'apparition d'une forme permet d'obtenir des concordances classiques où le mot est mis en évidence et inséré dans ses différents contextes, les uns à la suite des autres.

[2] Nous utilisons une fonction rémanente de calcul du logarithme népérien du coefficient du binôme, mémorisant les appels de (1,1) à (100,100), reposant sur une fonction rémanente de calcul du logarithme népérien de la factorielle, mémorisant les appels de 1 à 100, utilisant elle-même la fonction gammaln de la librairie NRC (Press W.H. *et al.*, 1992).

3.2. *Concordances KWIC*

Systématisons plus avant la lecture des contextes. Partant des constats que :

– la lecture des contextes d'apparition d'une notion procède le plus souvent d'une lecture allant de la notion vers le début ou vers la fin du contexte ;
– des contextes d'apparition similaires ont tendance à provoquer une catégorisation similaire de la notion analysée (nous assimilons ici le travail de dépouillement des contextes d'apparition d'une base à celui de la catégorisation des notions qu'elle sous-tend).

Une variante de l'outil *Contexte*, appelé *Concordances*, nous permet de synthétiser de manière plus efficace encore les contextes d'apparition. L'outil Concordances ne se distingue des Contextes que par seulement deux points, mais essentiels :

– chaque contexte est affiché sur une seule ligne. Ceci permet d'aligner les apparitions de la notion les unes au dessus des autres ;
– les lignes sont triées selon le contexte gauche ou droit en fonction de la notion étudiée et notamment des propriétés morphosyntaxiques des mots qui la réalisent.

Dans ces conditions, un usage approprié de tris multiples permet d'obtenir une liste de contextes qui *rapproche* les apparitions situées dans des contextes similaires. Ce rapprochement est alors utilisé comme une heuristique de lecture. La figure 2 présente un extrait de la concordance KWIC de Constitution chez Mirabeau triée à droite.

MIR26, p812	peuples . on dénonce de toute part la	**Constitution**	civile du clergé , décrétée par vos
MIR25, p810	que l' exposition des principes de la	**Constitution**	civile du clergé , récemment publiée
MIR25, p798	à ce que vous avez statué sur la	**Constitution**	civile du clergé ; mais que vous
MIR26, p829	à sceller de votre serment la nouvelle	**Constitution**	civile du clergé que par l'
MIR30, p850	serait donc pas une , surtout dans une	**Constitution**	comme la nôtre , dont le premier
MIR26, p829	entraîner dans sa chute la liberté et la	**Constitution**	de l' empire . l' une n' aspire à voir
MIR25, p805	oubli des principes élémentaires de la	**Constitution**	de l' Église . sans rechercher en quoi
MIR20, p747	n' avions pas eu le droit de changer la	**Constitution**	de l' État , ou que l' exercice d
MIR25, p803	sacerdotales . on cherche à paralyser la	**Constitution**	de l' État , pour faire revivre l'

Figure 2.
Extrait de la Concordance KWIC de « Constitution » chez Mirabeau triée à droite. Les contextes sont volontairement réduits pour des raisons de place, la référence renvoie à la page intégrale de l'édition pour la lecture élargie.

Au début de chaque contexte, la zone soulignée en bleu forme la *référence* qui situe l'apparition dans l'œuvre et donc dans le corpus. Elle est composée d'une réduction du nom de l'orateur (MIR), du n° de son discours (01, 03…), puis du numéro de page. La référence renvoie, de plus, par le biais d'un lien hypertextuel, directement à la page correspondante de l'édition en ligne du corpus des orateurs. A l'aide d'un simple lien, on accède donc aisément au contexte élargi de Constitution à travers la page qui contient l'occurrence du mot, tout en pouvant revenir aussi facilement aux Contextes d'où on est arrivé. La figure 3 présente un exemple de page d'édition en ligne, la page 805 de l'œuvre. La page d'édition est conçue de sorte à représenter le plus fidèlement possible le fac-similé de l'œuvre d'origine à l'aide de la mise en page, de la typographie, etc. Elle est elle-même bien sûr reliée aux autres pages de

l'œuvre par des liens hypertextuels. Ce réseau de pages constitue de fait le contexte définitif de la réalisation du champ sémantique, c'est-à-dire la lecture de l'ensemble de l'œuvre. On notera que le lien hypertextuel de la référence relativise la gêne occasionnée par la taille limitée du contexte affiché (même si la taille de ce contexte est paramétrable et volontairement limité ici) : le rôle du contexte est, en quelque sorte, de permettre une présélection focalisée sur un champ, pour approfondir, éventuellement, vers les pages de lecture complètes. On peut, en ce sens, parler d'un premier niveau de lecture « dynamique » focalisée, dont le support est un hypertexte.

Cette concordance est triée selon le contexte droit. L'ordre lexicographique des contextes droits à permis « d'empiler » les apparitions de Constitution participant aux mêmes locutions comme « Constitution civile du clergé » ou « Constitution de l'État ». Ici, le tri de concordances nous a permis de regrouper ensemble pour l'analyse les locutions dans lesquelles la base participe, locutions que l'outil d'analyse du vocabulaire n'avait pas construit initialement. Dans ce cas, on peut alors focaliser l'étude de la notion à partir de la locution elle-même, voire mettre à jour le corpus en y forçant cette locution comme unité lexicale, de sorte à ce qu'elle fasse partie de son vocabulaire de base.

Le réglage des divers tris ainsi que le parcours hypertextuel de l'édition du corpus nous font ici entrer dans un usage dynamique de l'outil lexicométrique où on ne peut plus se contenter de dépouiller des listings imprimés sur le papier[3]. La paramétrisation de l'instrument donne accès au corpus selon divers prismes et rend la lecture de ce dernier dynamique : ce qu'on y voit ou trouve dépend des réglages des parcours transversaux que l'on y effectue. En opposition au dépouillement d'un listing, ceci permet d'engager une sorte d' « interaction » avec le corpus.

En consultant la colonne des références de cette concordance (la première colonne), on peut par ailleurs constater que l'ordre de présentation des apparitions de Constitution ne correspond plus à l'ordre naturel du texte du corpus. De fait, cette délinéarisation du texte, provoquée par le tri, entraîne une lecture paradigmatique du matériau textuel. Et c'est ce type de lecture que nous allons continuer à suivre dans la suite de cet article. Bien sûr le lien associé à la référence (soulignée en bleu) donne immédiatement accès à la page où se trouve l'occurrence, ce qui permet toujours de revenir à la linéarité « naturelle » du texte.

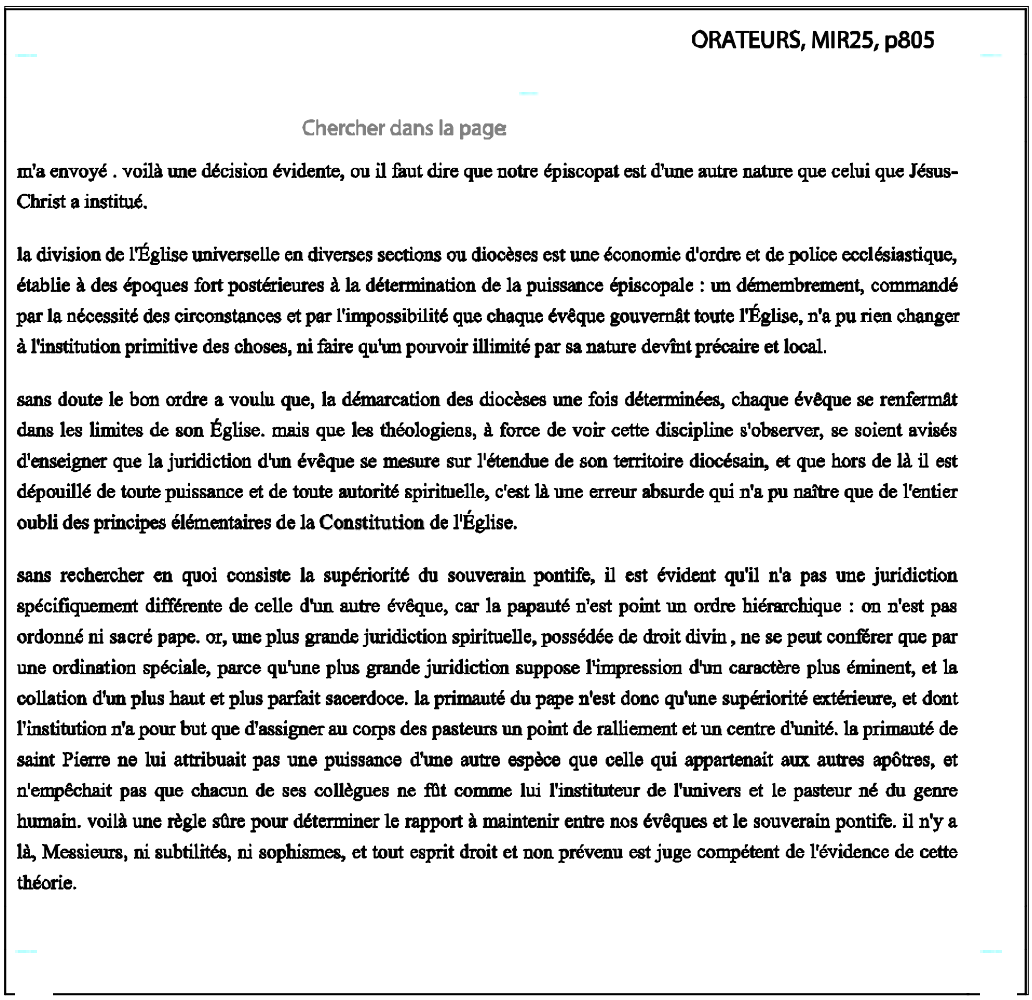
La concordance KWIC triée forme le deuxième niveau de synthèse paradigmatique de *Weblex* après le premier qui est lui constitué par les contextes simples d'apparition.

3.3. *Le lexicogramme de « Constitution » chez Mirabeau*

Un des problèmes des concordances KWIC triées est que seuls certains rapprochements de fonctionnements discursifs contigus, comme dans les locutions, sont offerts immédiatement à la lecture. Or, de nombreux liens de cooccurrence « à distance » sont susceptibles d'intéresser un dépouillement notionnel. L'outil *Lexicogramme* va être un moyen de palier à ce problème à l'aide de l'indice quantitatif de cooccurrence présenté à la section 2. Le lexicogramme d'un mot s'interprète comme une synthèse des cooccurrents gauches et droits d'un mot, à l'intérieur de toutes les phrases où il apparaît. Il peut aussi s'interpréter approximativement comme les listes hiérarchiques du vocabulaire des contextes gauches et droits d'une concordance KWIC. Le mot faisant l'objet du lexicogramme est appelé pivot du lexicogramme. Afin

[3] Ceci n'enlève rien au confort naturel de la lecture sur le papier, qui reste nécessaire à certains moments du dépouillement quand les données restent en quantités importantes.

d'illustrer l'usage de cet outil, et dans la continuité de l'effort de synthèse de sa concordance KWIC triée, la figure 4 présente le lexicogramme de Constitution chez Mirabeau. Les colonnes de gauche renseignent sur les cooccurrents situés à gauche de Convention dans le texte (en probabilité), les colonnes de droite sur les cooccurrents situés à droite.


ORATEURS, MIR25, p805

Chercher dans la page

m'a envoyé . voilà une décision évidente, ou il faut dire que notre épiscopat est d'une autre nature que celui que Jésus-Christ a institué.

la division de l'Église universelle en diverses sections ou diocèses est une économie d'ordre et de police ecclésiastique, établie à des époques fort postérieures à la détermination de la puissance épiscopale : un démembrement, commandé par la nécessité des circonstances et par l'impossibilité que chaque évêque gouvernât toute l'Église, n'a pu rien changer à l'institution primitive des choses, ni faire qu'un pouvoir illimité par sa nature devînt précaire et local.

sans doute le bon ordre a voulu que, la démarcation des diocèses une fois déterminées, chaque évêque se renfermât dans les limites de son Église. mais que les théologiens, à force de voir cette discipline s'observer, se soient avisés d'enseigner que la juridiction d'un évêque se mesure sur l'étendue de son territoire diocésain, et que hors de là il est dépouillé de toute puissance et de toute autorité spirituelle, c'est là une erreur absurde qui n'a pu naître que de l'entier oubli des principes élémentaires de la Constitution de l'Église.

sans rechercher en quoi consiste la supériorité du souverain pontife, il est évident qu'il n'a pas une juridiction spécifiquement différente de celle d'un autre évêque, car la papauté n'est point un ordre hiérarchique : on n'est pas ordonné ni sacré pape. or, une plus grande juridiction spirituelle, possédée de droit divin , ne se peut conférer que par une ordination spéciale, parce qu'une plus grande juridiction suppose l'impression d'un caractère plus éminent, et la collation d'un plus haut et plus parfait sacerdoce. la primauté du pape n'est donc qu'une supériorité extérieure, et dont l'institution n'a pour but que d'assigner au corps des pasteurs un point de ralliement et un centre d'unité. la primauté de saint Pierre ne lui attribuait pas une puissance d'une autre espèce que celle qui appartenait aux autres apôtres, et n'empêchait pas que chacun de ses collègues ne fût comme lui l'instituteur de l'univers et le pasteur né du genre humain. voilà une règle sûre pour déterminer le rapport à maintenir entre nos évêques et le souverain pontife. il n'y a là, Messieurs, ni subtilités, ni sophismes, et tout esprit droit et non prévenu est juge compétent de l'évidence de cette théorie.

Figure 3. Édition en ligne de la page 805 du corpus des orateurs. Cette page se situe dans le discours n°25 et fait partie d'un discours de Mirabeau comme l'indique son en-tête. Le mot Constitution y est mis en évidence en couleur rouge car la page a été accédée à partir d'un lien hypertextuel issu d'une concordance de ce mot. Les flèches situées aux quatre coins de la page sont des liens hypertextuels vers les pages précédentes et suivantes.

En fait la liste des cooccurrents gauches, par exemple, d'un pivot est potentiellement l'ensemble de tout le vocabulaire se trouvant à sa gauche dans les phrases, qu'ils lui soient contigus ou non. Chez Mirabeau il s'agit de 747 mots différents situés à gauche de Constitution dans ses discours. Afin d'obtenir une liste exploitable (ou lisible), c'est-à-dire plus limitée et constituée des seuls mots « les plus cooccurrents avec » ou « les plus attirés par » Constitution, nous utilisons le modèle probabiliste de cooccurrence. Ce modèle nous permet de trier la liste des mots cooccurrents afin de pouvoir n'afficher que ses premiers éléments, et

c'est sa seule vocation[4]. Un paramétrage de seuils permet alors de faire varier le nombre maximum de mots cooccurrents que l'on désire afficher. Dans l'usage de ce modèle, le chercheur a donc un rôle actif de réglages de l'instrument d'analyse. En aucun cas il s'agit d'essayer d'interpréter une « réalité » sous-jacente calculée par la machine, mais plutôt d'opérer un parcours interprétatif en « filtrant » à la demande la richesse de l'espace de cooccurrence du corpus utilisé.

Les lexicogrammes peuvent être triés selon leurs différentes colonnes afin d'orienter la lecture. Les tris et les seuils les plus utilisés sont ceux en probabilité de cooccurrence et en distance moyenne (dont le calcul est tout à fait indépendant de celui de la probabilité de cooccurrence). Dans la lecture du lexicogramme ces deux dimensions sont utilisées conjointement pour interpréter le lien de cooccurrence : les attirances fortes de mots rapprochés, en moyenne, correspondent aux figements lexicaux, aux locutions, voire aux syntagmes, les attirances fortes de mots plus éloignés, correspondent plus aux fonctionnements discursifs, voire thématiques des cooccurrents. Enfin, comme la lecture des lexicogrammes, qui forment une sorte de synthèse de la contextualisation de leur pivot — soit une synthèse de concordance KWIC — emmène souvent la lecture des propres lexicogrammes des cooccurrents du pivot courant, *Weblex* fournit un lien hypertextuel direct vers le calcul du lexicogramme de chaque cooccurrent à travers le clic sur sa forme.

Les lexicogrammes forment le troisième niveau de synthèse paradigmatique de *Weblex*.

Constitution (153)									
cooccurrents gauches					**cooccurrents droits**				
	f	**cf**	**p**	**d_m**		**f**	**cf**	**p**	**d_m**
comité	32	8	1e-05	1.0	consacrés	7	4	5e-05	6.8
Déclaration	16	6	1e-05	8.8	gouvernement	45	7	9e-04	15.6
principes	124	14	8e-05	7.3	française	23	5	1e-03	1.6
royal	5	3	4e-04	14.0	principes	124	11	4e-03	15.2
nouvelle	29	6	4e-04	0.0	résistance	19	4	4e-03	7.0
concilier	9	3	3e-03	4.0	civile	21	4	6e-03	0.0
changer	18	4	3e-03	10.2	voeux	11	3	6e-03	8.0
rapport	30	5	4e-03	6.2	délégués	12	3	8e-03	6.0
rédaction	11	3	6e-03	13.0	désormais	12	3	8e-03	23.3
organisation	13	3	1e-02	14.3	maintenir	13	3	1e-02	7.7
esprit	39	5	1e-02	4.4	exécution	15	3	1e-02	17.7
ancienne	14	3	1e-02	7.7	matière	17	3	2e-02	32.0
droits	109	9	1e-02	6.7	entièrement	19	3	3e-02	2.7
veto	41	5	1e-02	14.2	égalité	19	3	3e-02	5.3
voir	28	4	2e-02	32.0	État	67	6	3e-02	11.7
doit	133	10	2e-02	12.5	rendre	34	4	3e-02	5.2
lui-même	32	4	3e-02	11.2	jour	34	4	3e-02	15.5
arrêter	20	3	3e-02	13.0	part	21	3	4e-02	29.7
travail	21	3	4e-02	17.0	social	22	3	4e-02	9.3

Figure 4. Lexicogramme de « Constitution » chez Mirabeau.

*A droite des colonnes de formes de cooccurrents gauches et droits, la colonne **f** donne la fréquence du cooccurrent, **cf** la co-fréquence ou nombre de rencontres avec le pivot dans les phrases du corpus, **p** la probabilité de cooccurrence calculée et **d_m** la distance moyenne, en*

[4] Un modèle de cooccurrences en discours « complet » devrait au moins aussi tenir compte d'attirances distributionnelles en langue, ce qui n'est pas le cas ici.

*nombre de mots, séparant le cooccurrent du pivot dans le corpus. Pour afficher ce lexicogramme, les seuils : **f** ≥ 3, **cf** ≥ 3, **p** ≤ 5.0E-2, d_m ≤ 1000.0, ont été utilisés.*

3.4. *Comparaison entre lexicogrammes*

Ces synthèses de cooccurrents peuvent bien sûr se lire en comparaison les unes avec les autres. La figure 5 présente ainsi le lexicogramme de Constitution chez Sieyès. On accède alors de manière très synthétique à la réalisation de ce champ chez ces deux orateurs.

Constitution (21)

cooccurrents gauches					cooccurrents droits				
	f	**cf**	**p**	**d_m**		**f**	**cf**	**p**	**d_m**
raisonnée	4	3	9e-05	25.0	constituant	10	2	3e-02	13.0
exposition	4	3	9e-05	26.0	appartient	11	2	4e-02	2.5
réformer	5	3	2e-04	11.7	présenter	11	2	4e-02	16.5
bonne	5	2	8e-03	0.0	donner	13	2	5e-02	5.5
française	6	2	1e-02	17.0	objet	15	2	7e-02	22.0
parties	12	2	5e-02	2.0	publics	16	2	8e-02	8.5
partie	17	2	9e-02	14.5	peuple	25	2	2e-01	7.5
droits	47	3	1e-01	26.3	pouvoirs	26	2	2e-01	7.5
nation	48	3	1e-01	16.0	pouvoir	61	3	2e-01	9.7
moyens	24	2	2e-01	8.5					
citoyen	24	2	2e-01	27.0					
peuple	25	2	2e-01	9.0					
homme	30	2	2e-01	30.0					
doit	44	2	4e-01	7.0					

Figure 5. Lexicogramme du pôle « Constitution » dans les discours de Sieyès.
*Seuils : **f** 3, **cf** 2, **p** 5.0E-1, d_m 1000.0*

3.5. *Descente de contrôle vers les concordances de couples*

L'interprétation complète (ou fine) d'un couple de cooccurrents donné, dépend de la lecture précise de leurs contextes de rencontre. L'outil *Weblex* fournit donc un lien hypertextuel (associé à la fréquence **cf** de leur rencontre, soulignée en bleu, dans les lexicogrammes) provoquant le calcul de la concordance KWIC de l'apparition effective du couple dans le corpus. La figure 6 illustre, en exemple, la concordance obtenue en cliquant sur la co-fréquence « 7 » de la deuxième ligne des cooccurrents droits du lexicogramme de Constitution (voir la figure 4), c'est à dire le lien vers la concordance du couple (Constitution – gouvernement).

L'accès à ces concordances permet de « descendre » d'un niveau de synthèse paradigmatique, les concordances donnant elles-mêmes accès au niveau de lecture totale du corpus qui forme lui-même le niveau de base. Un principe fondamental du parcours hypertextuel offert par l'outil *Weblex* est donc le suivant : à chaque **montée en synthèse** paradigmatique doit correspondre une possibilité de **descente de contrôle** vers le niveau de synthèse inférieur.

1	MIR11, p664	politique a le droit inaliéna-ble d' établir , de modifier ou de changer la	**Constitution , c' est-à-dire la forme de son gouverne-ment**	, la distribution et les bornes des différents pouvoirs qui le composent .
2	MIR11, p666	des grands États , et surtout de l' empire français , que chaque progrès dans leur	**Constitution , dans leurs lois , dans leur gouverne-ment**	, agrandit la raison et la per-fectibilité humaine . elle vous sera due , cette
3	MIR20, p754	ême mon profond regret , que l' homme qui a posé les bases de la	**Constitution , et qui a le plus contribué à votre grand ouvrage , que l' homme qui a révélé au monde les véritables prin-cipes du gouvernement**	représentatif , se condamne lui-même à un silence que je déplore , que je trouve c

Figure 6.

Trois premières lignes de Concordance des sept rencontres de « Constitution » suivi de « gouvernement » dans les discours de Mirabeau. Comme pour les concordances précédentes, la référence de la ligne de concordance permet d'accéder à la page d'apparition de l'occurrence du couple pour une lecture plus approfondie dans l'archive elle-même.

4. Conclusion : le lexicogramme récursif de « Constitution » chez Mirabeau et Sieyès

Le parcours successif, de cooccurrents de cooccurrents, etc, à travers le réseau de lexicogrammes, permet d'accéder à une certaine image synthétique de plus en plus raffinée de la contextualisation d'un pivot initial. *Weblex*, en synthèse supérieure, permet d'afficher l'image de la totalité de ce parcours sous la forme d'un graphe appelé lexicogramme récursif. Pour construire ce graphe, l'outil parcours lui-même l'ensemble des lexicogrammes jusqu'à saturation du vocabulaire, puis dessine le graphe correspondant au parcours. Bien sûr aux seuils de calcul des lexicogrammes calculés précédemment, le graphe de parcours serait trop grand pour être représenté sur une page. L'outil cherche donc automatiquement un seuil en probabilité de sorte à obtenir un graphe ayant un nombre maximum prédéfini de mots cooccurrents. De plus, un seuil supplémentaire **pl** (pour palier) est utilisé afin de limiter la profondeur du parcours à partir du pivot. Dans un lexicogramme récursif, chaque nœud représente une forme du vocabulaire (présente une seule fois dans le graphe par définition), et chaque arc un lien de cooccurrence entre les nœuds où l'étiquette indique la force de la cooccurrence[5]. La figure 7 présente le lexicogramme récursif de Constitution chez Mirabeau, puis la figure 8 celui de Constitution chez Sieyès. Dans le même esprit de contrôle du graphe de cooccurrence obtenu, *Weblex* associe un lien hypertexte à chaque nœud du graphe vers le calcul de son lexicogramme (plus détaillé), donnant lui-même accès aux concordances de couples de cooccurrents, elles mêmes donnant accès aux pages d'édition où ces couples apparaissent.

L'implémentation de la méthode lexicométrique dans l'outil *Weblex* utilise donc la métaphore de l'hypertexte pour favoriser le va-et-vient nécessaire entre la montée en synthèse assistée par des indices quantitatifs et les descentes de contrôle dans la colonne paradigmatique d'un corpus donné. C'est la dynamique du parcours qui construit l'interprétation et non la lecture de résultats statiques donnés une fois pour toute. Cette dynamique est contrôlée par le réglage de paramètres variés : le choix de tri des cooccurrents, les seuils d'élagage quantitatifs (sur le

[5] Précisément : l'étiquette correspond au logarithme de la probabilité de cooccurrence entre les nœuds, plus le nombre est grand plus la probabilité de rencontre est faible, et donc plus l'étonnement est grand et le couple cooccurrent.

modèle comme la valeur de probabilité maximale à considérer, et sur l'interface comme le nombre maximum d'éléments à lister, le nombre maximum de nœuds d'un graphe, son algorithme de placement), les moyens d'élagage qualitatifs (prise en compte ou non de mots particuliers, de mots outils, numéraires, etc.), le choix du contexte de rencontre, les choix initiaux de segmentation des lexies (lexèmes et/ou locutions). Nos développements futurs concernent d'une part une meilleure prise en charge de réglages semi-automatiques de parcours, notamment au passage entre les différents niveaux de synthèse où les seuils d'élagage (par exemple) peuvent être adaptés semi-automatiquement en fonction de la quantité d'informations à afficher, et d'autre part la possibilité de déposer des annotations au fil d'un parcours afin de pouvoir mémoriser la progression d'une interprétation et éventuellement de la transmettre pour évaluation.

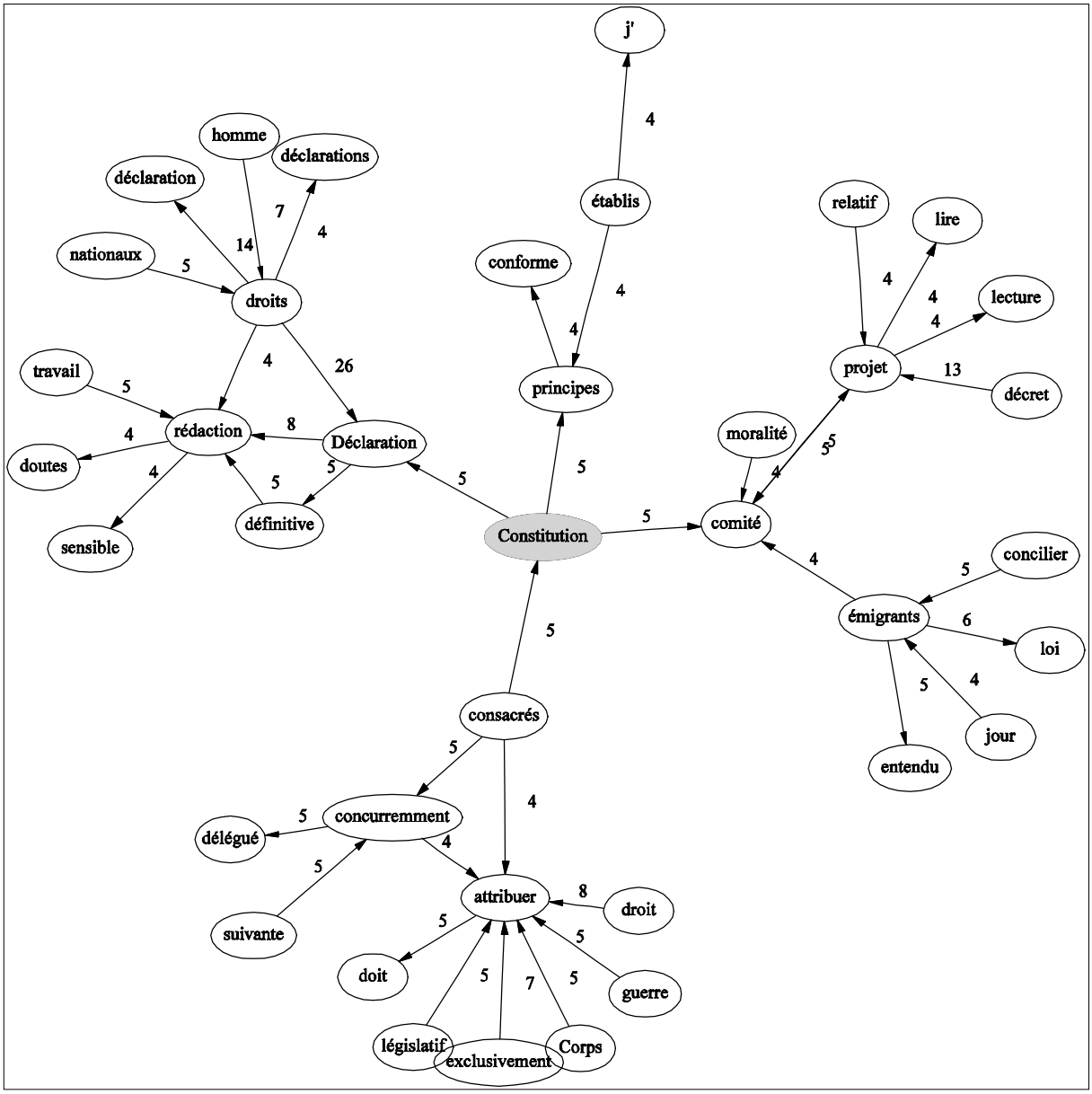


*Figure 7. Lexicogramme récursif du pôle « Constitution » dans les discours de Mirabeau. Seuils : **p** 4e-04, **r** 2, **f** 3, $\boldsymbol{d_m}$ 1000.0, **pl** 3*

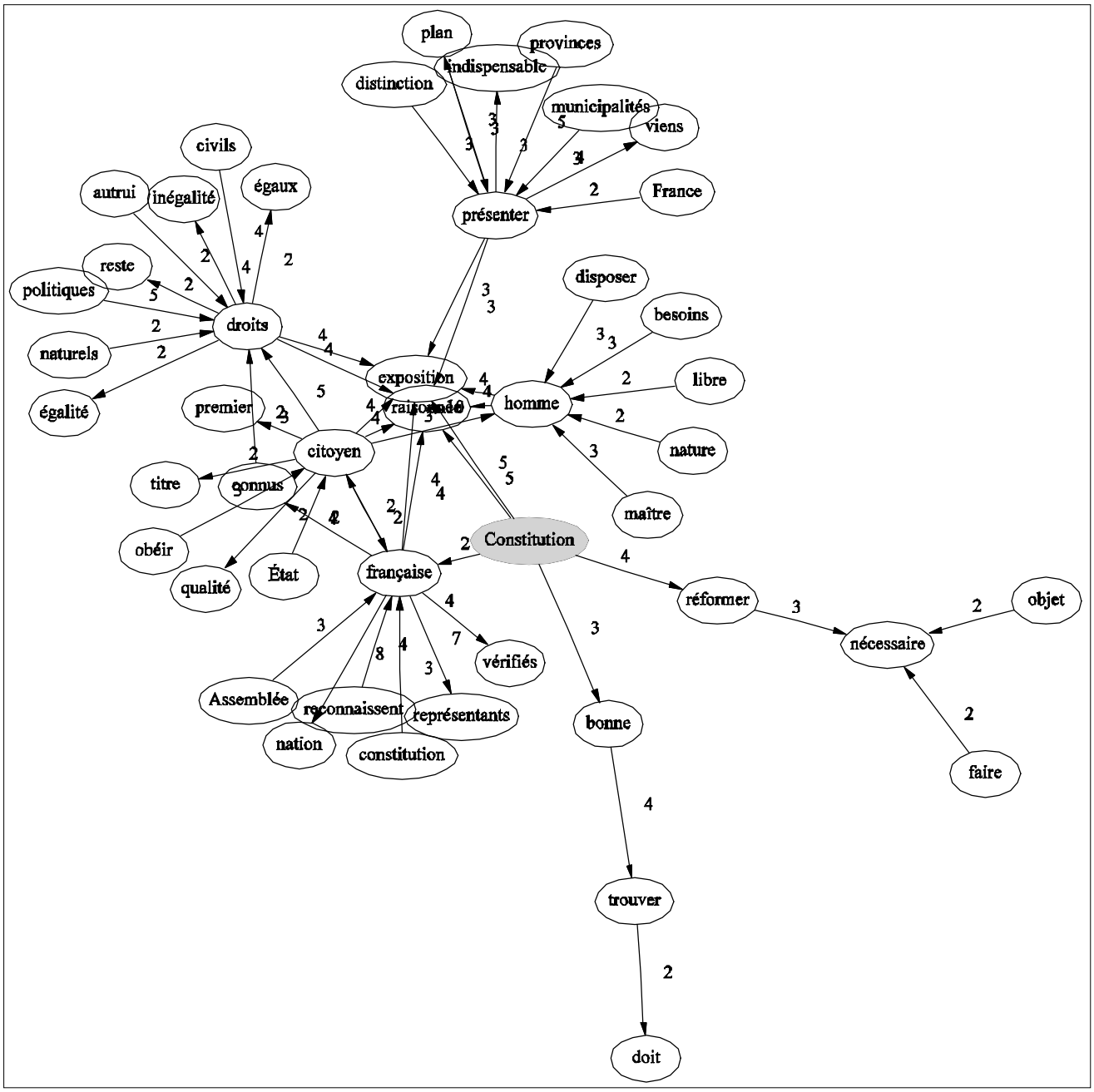


Figure 8. Lexicogramme récursif du pôle « Constitution » dans les discours de Sieyès.
*Seuils : **p** 2e-02, **r** 2, **f** 3, $\mathbf{d_m}$ 1000.0, **pl** 3*

Références

Heiden S. (2002). *Manuel Utilisateur de Weblex. Version 4.1*. Janvier 2002, ICAR CNRS/ENS-LSH, <http://weblex.ens-lsh.fr/doc/weblex.pdf>.

Krenn B. et Evert S. (2001). Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of ACL Workshop on Collocations*, Toulouse.

Labbé D. et Labbé C. (2001). Inter-Textual Distance and Authorship Attribution. Corneille and Molière. *Journal of Quantitative Linguistics*, vol. (8) : 213-231.

Lafon P. (1984). *Dépouillements et Statistiques en Lexicométrie*. Slatkine-Champion.

Manning C.D. et Schütze H. (2002). *Foundations of statistical natural language processing*. MIT Press : 151-189.

Oakes M.P. (1998). *Statistics for Corpus Linguistics*. Edinburgh University Press.

Press W.H. *et al.* (1992). *Numerical Recipes in C : The Art of Scientific Computing*. Cambridge University Press.

Mapping the structure of research topics through term variant clustering: the *TermWatch* system

Fidelia Ibekwe-SanJuan[1], Eric SanJuan[2]

[1] ERSICOM
Université Jean Moulin – 4 cours Albert Thomas – 69008 Lyon.
ibekwe@univ-lyon3.fr

[2] LITA EA3097 – Université de Metz.
Île de Saulcy – 57047 Metz – France.
eric.sanjuan@iut.univ-metz.fr

Abstract

A multi-disciplinary approach integrating computational linguistic techniques is necessary to elaborate indicators of research topic evolution. We describe a system which bases clustering on linguistic relations, instead of the usual co-occurrence paradigm. The interesting features of this approach, embodied in the *TermWatch* system, lie in the combination of state-of-the-art techniques in computational terminology, mathematics (graph formalism) and visualization techniques. Computational terminology enable us to extract meaningful text chunks and to relate these chunks through linguistic relations. These text chunks are terms and the linguistic relations are syntactic variations. We integrated into this system an adapted visualization tool which enhances comprehension of the research topic layout and their trends. Here we focus on the chronological analysis of graphs issued by *TermWatch* through a graph visualization tool, *Aisee* which helps the end-user to track the main tendencies of research topics in his/her field.

Résumé

Nous présentons les dernières avancées apportées au système *TermWatch*, élaboré à des fins de veille scientifique et technologique ou de fouille de textes. La principale originalité de *TermWatch* réside dans le fait que la classification est fondée sur des relations syntaxiques, et non sur des critères usuels de co-occurrence. Une interface de visualisation, *Aisee* a été récemment intégrée au système pour mieux explorer ses résultats. Cet article met l'accent sur l'analyse des résultats à travers cette interface, notamment l'analyse chronologique permettant de percevoir l'évolution des tendances thématiques dans un corpus de textes spécialisés.

Keywords: Morpho-syntactic analysis, terminological variation, clustering, textmining, trend mapping, competitive intelligence.

1. Introduction

Bibliometric and scientometric studies aim to elaborate indicators of scientific activities through the use of statistical data analysis and mathematical models. The two major methods used in these fields are the co-citation (Small, 1999) and co-word analyses (Callon *et al*., 1983). While citation analysis has proved its usefulness for highlighting major actors in a field (the "who's who" of a field), we argue that it is not adequate to portray the publication contents themselves and their evolution. It cannot capture other facets of scientific activities. For instance, how can one know actually what themes or topics are addressed in the different publications, if the research topics are evolving and what their relation to one another is. Still, citation studies have to say something about the contents of the publications in order to interpret the major citation trends.

Most bibliometric and scientometric methods work mostly at the macro level, i.e., the level of whole disciplines (chemistry, physics, mathematics, linguistics, etc), or whole countries or continents. Consequently, their data are characterized by very highly occurring units where low occurring units or rare phenomena tend to be simply eliminated. A lot of work has been devoted lately to text data analysis (Lebart and Salem, 1994; Reinert, 1993; Lelu, 2001, François *et al.*, 2001). While some of these works perform some linguistic processing of the text prior to the clustering, the unique criterion for clustering in these data analysis methods remain the co-occurrence paradigm. We claim that since the information units being clustered come from texts written in natural language, related on other more linguistic dimensions (morphological, lexical, syntactic, semantic), it is necessary to explore which linguistic relations are potentially relevant as clustering criteria. The system we developed, *TermWatch* is based on this claim. It clusters terms extracted from texts, based on syntactic relations called variations. This system is more adapted to capturing rare occurring local phenomena as well as highly occurring ones since clustering is not based on frequency. Low and highly occurring units are given the same chance by our method. Different stages of this work have been published elsewhere (SanJuan and Ibekwe-SanJuan, 2002; Ibekwe-SanJuan and SanJuan, 2003). The focus of this paper is on the latest enhancements to *TermWatch:* the integration of a graph visualization tool, *Aisee*[1], for exploring the results and for performing a chronological analysis in order to detect trends.

First, we present the overall architecture of the *TermWatch* system (§2) and describe it's different levels of processing – term extraction from texts followed by terminological variation identification and finally by clustering. In section §3, we present the *Aisee* interface and illustrate how it brings to light the organization of research topics in a field. We will then focus on the chronological analysis of the structure of research topics with the aim to pinpoint their evolution patterns. Finally, section §4 will be dedicated to discussions. The experimental corpus used in this study was composed of English scientific abstracts of 70,000 words collected from two scientific databases following a STW request. They covered publication made from 1988 to 1998 on the breadmaking process. The end user wished to know if there existed new natural additives to enhance his bread making process while maintaining it's "artisanal" quality. The corpus was made available by the French Institute for Scientific and Technical Information (INIST).

2. System overview

TermWatch comprises of two main modules: a term variant search module and a clustering module. Two other minor modules ensure the integration of two external tools necessary to perform the whole analysis: a term extractor and a graph visualization tool. It's architecture is shown in figure 1 here below. We recall briefly the different stages of processing involved.

2.1. *Term extraction with INTEX*

Given a corpus of english texts, the first step is to perform morpho-syntactic analysis in order to extract terms from the texts. We used the INTEX linguistic toolbox (Silberztein, 1993) for this task. Terms are choice linguistic units, rich in information content because they are used by experts in a field to name the objects or concepts of that particular field. On the linguistic level, terms appear mostly as noun phrases which can occur either in a compound form

[1] More details on this tool can be found at http://www.aisee.com

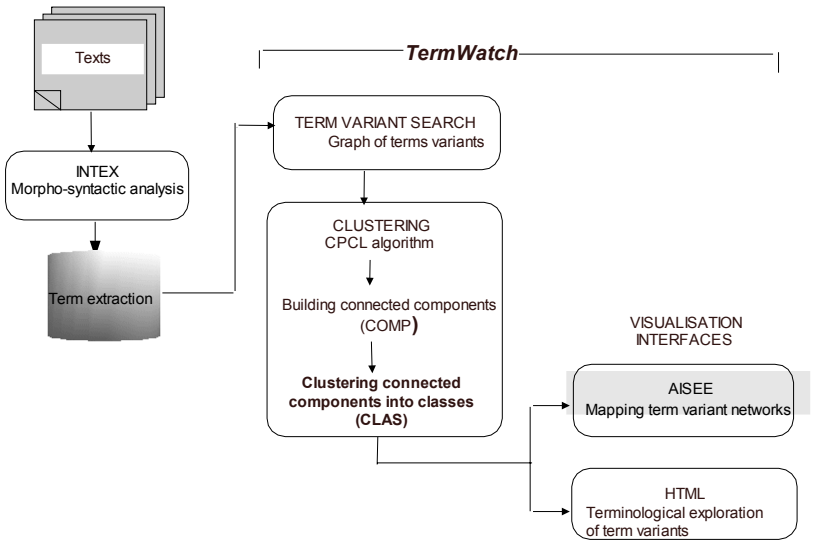


Figure 1. Overall system architecture

(*wheat flour fractionation*) or in a syntagmatic structure with a prepositional phrase (*fractionation of wheat flour*). After morphological analysis on the corpus, we defined several morpho-syntactic constraints enabling us to identify sequences that contained potential terms. We will not go into the details of the different grammar rules written to extract noun phrases. This stage has been described in SanJuan and Ibekwe-SanJuan (2002). We will simply say that these constraints are implemented as finite state transducers with decreasing order of complexity, some being embedded in others[2]. These transducers are applied in an iterative fashion on the corpus and enable us to extract first complex nominal sequences, which are in turn split into simpler noun phrases (NPs) until we reach the desired result. Examples of some candidates extracted are "*traditional sour dough starter cultures*; *development of traditional bread flavour*; *dough stickiness*; *intense dough stickiness*; *wheat dough stickiness*; *wheat dough surface stickiness*". Some 10 000 term candidates were thus extracted. In order to select likely domain terms, we subjected the list of NPs to manual filtering by a domain specialist (an indexer of the INIST used to indexing papers on this field). 3651 likely terms were retained. As terms are given to much variations in texts - the same term can be written in several ways, two terms will rarely co-occur under the same form. For instance, these variants around the term "*dough stickiness*" (*intense dough stickiness*; *wheat dough sickiness*; *wheat dough surface stickiness*) may only occur once in the corpus. This drives home the point that clustering methods based on co-occurrence of word sequences cannot work effectively on terms extracted from raw texts without addressing the terminological variation issue. *TermWatch* does not need any outside terminological resource for this term extraction task. This endogeneous approach is more likely to portray terminological evolution and hence that of domain concepts.

[2] A transducer is a graph whose vertices are morpho-syntactic tags. In the simple case, this graph is equivalent to a regular expression. It becomes a transducer when, in INTEX, it not only recognizes forms or patterns but also modifies the text.

2.2. *Syntactic variant search*

This module identifies pairs of terms that are related through some linguistic operations, thus making the two terms variants of one another. Variations occur at different linguistic levels making their identification impossible without integrating natural language processing (NLP) techniques. There are spelling variants (*specialisation* / *specialization*), morphological variants (*online web access* / *on line web access* / *on-line web access*; *WWW interface* / *web interface*), syntactic variants (*information retrieval* / *retrieval of information* / *efficient retrieval of information*) and finally semantic variants (*information retrieval* / *data access*). We did not address all the possible types of variants that can exist between terms (see Jacquemin, 2001 for an extensive study), rather we chose a restricted subset which can be identified with a rather shallow and selective NLP. These syntactic variants involved two categories of transformation: syntactic variants involving the addition of words in an existing term or the substitution of a word in an existing term. Within these two categories, we distinguished variants along the grammatical axis: variants affecting modifier words in a term and those affecting the head word. In the term "*traditional sour dough starter cultures*", the first four words are modifiers because they qualify the last noun "*cultures*" which is the head word. In other words, modifiers play the role of adjectives in a syntagm while the head word corresponds to the noun focus (the object of discourse). When considering terms with prepositional phrases, the head word appears as the last noun before the preposition as in "*effective retrieval of information*" where "*retrieval*" is the head word, the rest are its modifiers.

2.2.1. *Modifier variations (COMP)*

They involve the two transformations mentioned above: adjunction of words in an existing term (left expansions, insertions) and the substitution of a modifier word. For instance, "*new yeast strain development*" is an insertion (Ins) variant of "*new strain development*" while "*gas holding property of dough*" (=dough gas holding property) is a left expansion (L-Exp) of "*gas holding property*". "*commercial baking strain*" is a modifier substitution (M-Sub) variant of "*commercial yeast strain*". We call this category of variants that affect the modifier elements in a term "COMP".

2.2.2. *Head variations (CLAS)*

They equally subdivide into two operations: head expansions and head substitution. For instance, the term "*cell wall degrading enzyme*" is a head expansion (R-Exp) variant of "*cell wall*" where "enzyme" has become the new head of this syntagm. "*effect of xanthan gum*" is a head substitution (H-Sub) of "*addition of xanthan gum*". Left and right expansion (LR-Exp) represents a combination of the two elementary expansions as in "*bread dough*" and "*frozen bread dough preparation*". We call this category of relations "CLAS". The aim of the variant identification module is to identify such variants and establish a relation between each pair according to the type of phenomenon involved. 3019 terms (83%) out of the 3561 retained were thus related, showing the importance of the variation phenomena in written texts. Contrarily to other works on terminology variations (Daille, 2003; Jacquemin, 2001), we were not looking only for variants which preserved the semantic concept family of a given term, i.e; as in "*information retrieval*" and "*efficient information retrieval*", we would also extract as variants terms that share common elements as in "*information retrieval system*" and "*information retrieval software*", the thrust being on capturing what researchers are saying about a particular topic, in this example "*information retrieval*". As of now, the definition of terms and variations have been done only for the english language. However many studies already exist

on french term formation and variations. It will not require much effort to adapt these works in order to be able to process French texts.

2.3. *Term variant clustering*

This module is based on the CPCL (Classification Algorithm by Preferential Clustered Link) presented in Ibekwe-SanJuan (1998). Here, we extend its formal presentation. CPCL is a two-step extractor of classes from a graph of term variants implemented in *TermWatch*. One attribute of this algorithm is that the clustering begins not at the atomic level (term level), but at the component level. Components are obtained by grouping terms sharing COMP variations. The clustering stage then consists in merging iteratively components that share many variations of the type CLAS with regard to the links they share with any other component in the graph. A normalized coefficient is used to indicate the proximity between two components as a function of the number of CLAS relations between them and the proportion of the particular CLAS relation in the graph.

CPCL-phase 1: computing a dissimilarity index between components of terms

The first step of the CPCL algorithm is to compute a set *I* of components which are subsets of terms linked by variations of type COMP and a dissimilarity index between these components that will be used in the second step of the algorithm, which is a kind of single link clustering process.

The computing of the components is simply done by extracting the connected components of the cover graph of COMP. Thus a subset *S* of terms is a component if it is a maximal subset such that for any two terms t_0 and t_k in *S* there is a sequence of terms $t_1, \ldots, t_n$ such that (t_0,t_1), (t_1,t_2) $\ldots(t_{n-1},t_n)$, (t_n,t_k) are all linked by a relation in COMP. It follows that two terms in the same component share the same head word, but that two terms with the same head word are not necessarily in the same component if they do not share a COMP relation. Let us denote by *I* the set of these connected components. Like in most clustering methods, we need to compute a dissimilarity index.

By default, in the first release of *TermWatch*, CLAS is a set of four syntactic variation relations on terms: H-Sub-2 (head substitution on terms of length 2), H-Sub-3 (the same on terms with at least two modifiers), R-Exp (right expansion of terms), LR-Exp (left and right expansion). However, the user can chose to remove one of these relations if s/he finds it too noisy.

We denote by $n_R(i,j)$ the number of *R* variations between *i* and *j*, for any *i,j* in *I*. Clearly we have: $n_R(i,j) = |\{\{t,s\}: (t,s) \text{ in } R, t \text{ in } i, \text{ s in } j\}|$

Then for any *i,j* in *I*, we define a dissimilarity *d* that maps I^2 into [0,1] by setting: $d(i,j) = 1$ if we have $n_R(i,j) = 0$ for any R in CLAS ; $d(i,j) = 0$ whenever i = j ; otherwise:

$$d_{ij} = \sum_{R \in CLAS} \frac{n_R(i,j)}{|R|}$$

CPCL-phase2: Clustering components

The second step of the CPCL algorithm produces an appropriate visual display of complex information in CLAS in order to arrive at a better understanding of the network of term variants. To avoid the main drawback of single link clustering (SLC) called "chain effect", we have chosen to cluster first vertices that have the lowest dissimilarity by comparison with neighbouring vertices. In other words, we do not consider the values of *d* as an absolute

ordered set. We consider the relative strength of the link between any given pair of vertices at a given time. Thus, at a given iteration, two edges with different values can be clustered. In practice, this leads to a more fine-grained representation of the network of components related by CLAS links.

Using the CPCL-phase2 hierarchical cluster algorithm, dissimilarity *d* is represented by an ultrametric distance, called *lower differentiation ultrametric* (*ldu*), between the terminal nodes of a dendrogram. The ultrametric distance between two components is simply the level of the smallest class containing the two components. Each level of the dendrogram shows a possible classification of components, and consequently of terms. The *TermWatch* systems visualizes the significant levels as networks of classes using the *AiSee* graph display.

For any v in [0,1], let us denote by B(i,v) the set of pairs $\{x,y\}$ of components such that:

$$d(i,x) \leq v \text{ and } d(i,y) > v$$

and let us denote by $pr(v)$ the higher value w in the image of d such that $w< v$.

Then we have the following characterization of the *ldu.* It is smallest ultrametric u such that any pair i, j of distinct components we have:

1) if $u(i,j) = v <1$ then there exists a pair $\{x,y\}$ in B($i, pr(v)$) $\cap$ B($j, pr(v)$) such that:

$$d(x,y) = min\{d(z,w) : \{z,w\} \text{ in } B(i, pr(v)) \cup B(j, pr(v))\}.$$

2) $u(i,j) = min \{ d(x,y) : pr(u(x,y)) = pr(u(i,j)) \}$

Figure 2 is an example of the application of this algorithm to a dissimilarity on a small set represented by a valued graph. In this figure, circles show clusters obtained at the first iteration (first level of the dendrogram) and the triangles the two clusters formed at the second iteration. In this example the algorithm converges at the third iteration.

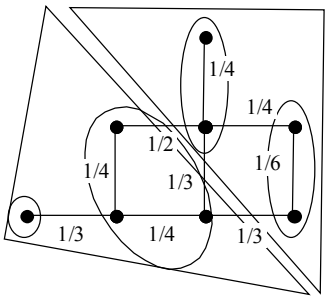


Figure 2. Running example of CPCL-phase2 algorithm

A comparison of the CPCL clustering algorithm with the single link clustering (SLC), done by Berry A, *et al.* (2004) shows it to produce different results that are more legible.

3. Visualization interface

The graphs produced by the clustering module represent classes that contain components connected by head variation relations (CLAS), these in turn contain sets of term variants connected by modifier variation relations (COMP). We posit that such graphs can depict the research topics present in the corpus. Hence, their mapping through a graphic display tool can enable the end user perceive the layout of research topics in his field. Through time series analysis, it can also enable him/her track the evolution of the topics (see §3.2 below). *TermWatch* generates undirected graphs whose layout is determined from the number of external

links between classes. Since a class has no coordinates, the space in which it is mapped is not geometric. A major difficulty for the visualization tool is to determine the position of each class such that its relation to other classes is not altered. Thus, it should allow the user to perceive the central classes and the peripheral ones. Also, such a tool should portray the notion of proximity even though we did not use any distance measure, i.e, the visualization tool should place classes that share strong variation links close together. Since the clustering method is based on variation relations, it is necessary that the visual tool allows the user to explore class contents up to the term level, in order to see the variants. Thus it should offer unfolding functions that ensure a three-level exploration of the classes: class, component and term level. It was also necessary to bring to light peculiar patterns, if any, formed by groups of classes sharing particular types of variation relations. The *Aisee* graph display package met all these requirements. It simulates the proximity of classes as edges with a spring embedder where close particles repel one another and distant particles attract each other. Integrating *Aisee* as the front end to *TermWatch* required specific encoding of *TermWatch's* output into the graph description language (GDL) used by *Aisee*.

The results obtained on the global corpus has been published elsewhere (Ibekwe-SanJuan and SanJuan, 2003), we will refer to this result as the "global classification". In the following sections, we will summarize the evaluation of this global classification, then focus on the results obtained by periods (through time series analysis) in order to pinpoint the evolution patterns of research topics.

3.1. *Evaluation of the global classification*

Thirty-three classes were obtained in the global classification at the 3rd iteration and were subjected to a scientific and technological watch (STW) analyst[3] for validation. The evaluation has been reported in Ibekwe-SanJuan and Dubois, (2002). The analyst had to say if a class represented a coherent domain topic, name the topic and also determine if the external links between classes were sound. In essence, the analyst's evaluation helped us identify three categories of classes: (a) classes that represented known and relevant domain topics, they were twenty-six in number); (b) classes whose topic were partially or not at all identified (five classes were concerned, especially the core class 32); and finally (c) classes whose topic though identifiable were uninteresting for STW, only one class is concerned. Two of the unidentified classes were rather big in size (198 and 218 terms respectively). This poses a problem for scientific and technological watch because the class contents were too heterogeneous to be qualified. Currently Berry *et al*. (2004) are investigating ways of splitting such big graphs in order to reduce them to more manageable and interpretable sub graphs.

Let's recall that our corpus covered ten years of publications (1988 to 1998) on the topic of bread production process. The initial STW request was to know if there existed new and natural additives that can conserve the "artisanal" quality of bread while enabling the company to expand (become more industrial and competitive). Among the relevant classes, one class labelled "*wheat bran*" represented an emerging topic at the time of corpus constitution (1998). According to the analyst, its content provide clues to answer the initial STW request. The analyst renamed this class "*Natural components or elements*". This name could not have been deduced from the lexical form of the term variants in this class. So this is a pragmatic domain knowledge whose linguistic utterance did not appear on the surface level as terms in our cor-

[3] The evaluation was done by the Technology survey unit of the Henri Tudor Research center in Luxembourg (Centre de Veille Technologique).

pus. A closer look at the term variants in this class however shows its vocabulary to be relatively specific. The variants around "*wheat germ*, *corn bran*" and "*bran incorporation, raisin incorporation*" appeared only in this class. The evaluation was done before the *Aisee* graphic interface was integrated. With the graphic interface, this type of information is enhanced. The emerging character of this class, at the time of corpus constitution, is supported by its external position. The *wheat bran* class was not in a central position and is linked to 2 other classes only.

3.2. *Tracking the evolution of research topics: a chronological analysis*

We will explore here the possibilities of capturing the evolution of research topics via the visual interface. The first problem in a chronological analysis is to determine significant time intervals. Usually in bibliometric studies, finding time periods is done by partitioning the corpus in such a way that an approximately equal number of units is obtained per period. Following this method, we determined three periods covering the ten years of the publications. The first period, P1 spanned publications from 1988-92, the second P2 covered publications from 1993-95 and the third period P3, from 1996-98. All three had approximately the same number of terms (± 1000 terms). We tried clustering term variants in each period but were rapidly confronted with a problem: by partitioning our terms into the time intervals specified above, the clustering yielded enormous classes as early as the 1st iteration, whose contents were too heterogeneous to be meaningful. The reason for the rapid aggregation of classes soon became clear: since we were working on only $^1/_3$ of the terms in each period, the variation links between terms from different periods were lost, so the terms in a particular period tended to aggregate earlier. We then abandoned this classical approach and opted for a different solution: we worked on the whole set of terms but at a particular point, we looked at links appearing in a particular period. We then mark links from each period using colour codes and other graphic symbols. Comparing this with the classes obtained on global classification, we were able to follow progressively the formation of their components in time.

In the graphic interface generated under *Aisee*, each vertex bears a sequential number given to the class and a label. The label is chosen automatically by *TermWatch* and is the most active term variant in the class. Vertices either represent connected components or classes. Classes are highlighted by a background colour whereas connected components that were not aggregated into a class have a colourless background. Components in the same class appear draped in the same background colour as the class's colour, ensuring an easy reading. Also, different colours and graphics are used to differentiate the links from different periods: in this example, links from period P1 are in dotted lines, those of P2 in dashed lines and those of P3 in straight lines. For the end user, the visualization tool offers a more interactive and dynamic interface. The user can select the periods whose links s/he wishes to see, the tool automatically reloads the graph accordingly.

The *Aisee* visualization tool highlighted important and meaningful features of the graph of term variants generated by *TermWatch*. These features concern the notion of "distance" between two classes and peculiar patterns formed by subgraphs of the network which we will describe below. The length of an edge has a straightforward meaning here. The longer the edge, the weaker the link between two classes and thus the further they are from one another. Thus, we obtain an image where distance between vertices are meaningful whereas our clustering algorithm does not use any distance measure.

Globally, the spatial organization of classes across the three periods exhibit a well known form in bibliometrics: the distribution model of "core" and "scatter" (see figure 3 hereafter). A core network is surrounded by a series of smaller isolated networks. In this core network, a number of components aggregated most of the links and maintained themselves in this position across the different time periods. Zooming in on the core of the network shows it to be structured around the following components: "262[4]-*wheat flour replacement*", "228-*wheat flour protein concn*", "339-*flour protein content*", "361-*wheat flour protein*", "313-*aqueous wheat flour phase*" "6-*wheat flour dough*". The last component maintained itself at the heart of the network throughout the three periods. Thus, the topic of "*wheat flour*" appears to structure research in this field during the years considered. We found the same modifiers shared by classes in the complete graph in the global classification.

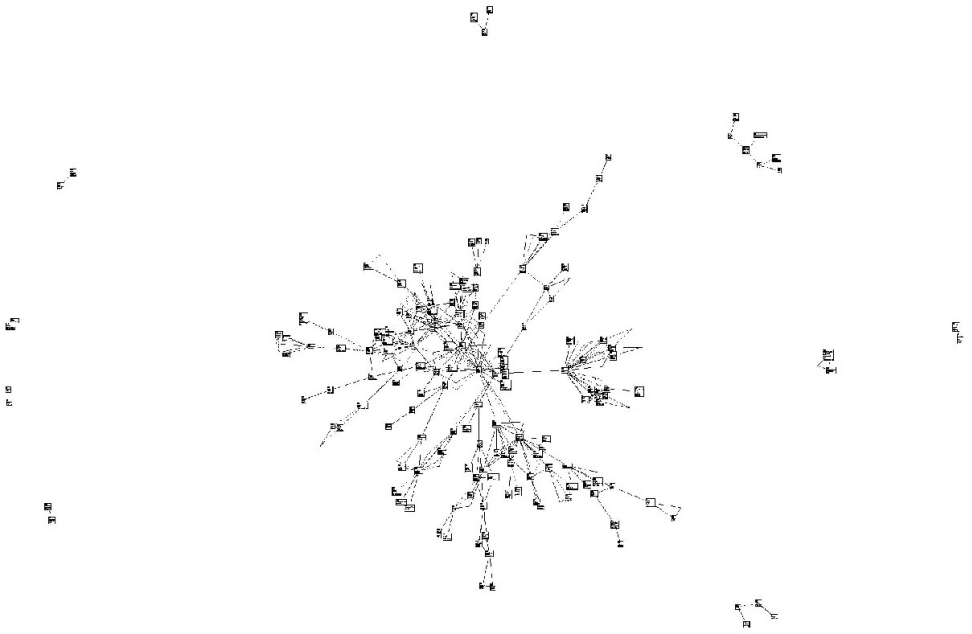


Figure 3. Global structure of the map across the different time periods

Some typical evolution patterns

Examining in more details the graphs we obtain as we add progressively links from different time periods, we observed four types of evolution patterns:

a- mergers between sub networks of components;

b- consolidation of sub networks of components by acquisitions of new components;

c- splitting of previously connected networks;

d- appearance or stability of little groups of isolated components.

Pattern (b) is exhibited by components composing the class whose content was deemed as "emerging" by the domain specialist (see §3.1). By introducing progressively links from different periods between components of this class, we were able to follow the formation of components in this class across the three time intervals. Figure 4 below show this movement.

[4] Number of the component.

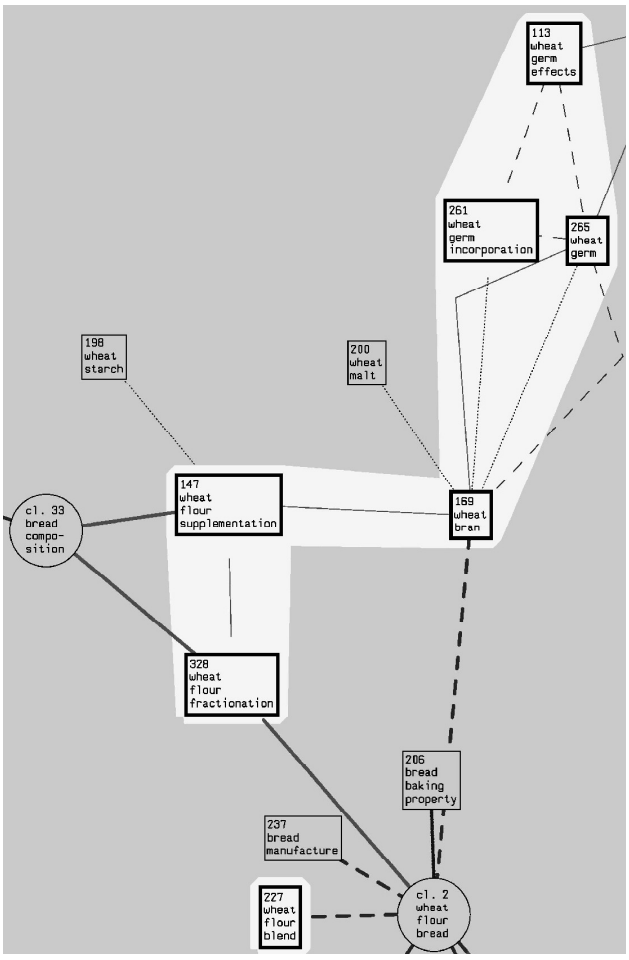


Figure 4. "Merger → consolidation" pattern exhibited by components of the emerging class in P3

We unfolded the class, so the image here below shows the components composing the class. This class contains seven components numbered by 169, 261, 265, 113, 227, 147 and 328.
The first components of this class appeared in period P1 and were represented by a little network around components "169-*wheat bran*" and "265-*wheat germ*" on the one hand, and components "147-*wheat flour supplementation*" and "169-*wheat bran*", in turn linked to component "*328-wheat flour fractionation*".

In period P2, a consolidation of this sub network is observed. Component "*169-wheat bran*" attaches itself to a central classes in the major network, "*2-wheat flour bread*", thus bringing the other components of this class in it's wake. At the same time, it consolidates its link with "*261-wheat germ incorporation*" with another link in period P2. New links between the three components 265, 261 and 113 appear, thus forming a complete graph.

In period P3, this consolidation movement is maintained by the acquisition of new links, between "169-*wheat bran*" and "265-*wheat germ*".

Since it is also possible to show links from different periods between components or classes, we are able to see that component "227-*wheat flour blend*" also belongs to this class owing to its links with components "147-*wheat flour supplementation*" and "328-*wheat flour fractionation*". For legibility reasons, links across two periods were not shown in the above figure but are highlighted by a different colour in the graphic interface. While this "appearance →

merger → consolidation" pattern brings this sub network in the vicinity of the main network, the components of the emerging class remain at the border of this central network, thus echoing the peripheral position of this class in the global classification and confirming it's potential as a weak but interesting signal for STW.

There are many more sub networks whose position and structure can be explored, but owing to space limitations, we cannot describe all of them in the present paper.

4. Discussion

The system we have described is designed for trends survey in a specific research field through the mapping of term variants extracted from a corpus of representative texts. We were not able to subject the chronological analysis to expert validation owing to the absence of a formal framework in which to carry out this evaluation. However, basing on the validation carried out on the global classification, the chronological analysis corroborated the general tendencies already observed on some of the classes: core *vs* peripheral, position of the emerging topic's class. The chronological analysis also brought to light the main evolution patterns of the classes.

We are currently investigating the extraction of semantic relations with which to enrich the clustering relations. This will enable us to establish equivalence relations between synonymic terms, thus ensuring that semantically related terms appear in the same class. As of now, such synonyms can be captured only if they share a common lexical element as in "*information retrieval*" and "*information access*" where "*information*" is the common element. These two will likely end up in the same class as they are considered head substitution variants, but so also would another head variant like "*information storage*". Although topically close, the latter is not exactly a synonym of the preceding two. However, its being in the same class with the former two is not necessarily a handicap for STW application. We also have to tackle other important morphological variants like abbreviations and acronyms which would also culminate in synonymy relations. Ways of identifying other semantic variants such as hypernym/hyponyms relations are currently under investigation.

References

Berry A., Kaba B., SanJuan E. and Sigayret A. (2004). Classification et désarticulation de graphes terminologiques. In *Actes des JADT 2004.*

Callon M., Courtial J-P. and Turner W. (1991). La méthode Leximappe: un outil pour l'analyse stratégique du développement scientifique et technique. In Vinck (Ed.), *Gestion de la recherche: nouveaux problèmes, nouveaux outils.* Boeck Editions: 207-277.

Daille B. (2003). Complex structuring through term variation. *Workshop on Multiword expressions: Analysis, Acquisition and Treatment.* In *41st Meeting of the Association for Computational Linguistics (ACL, 2003)*, Sapporo, Japan.

François C., Dubois C. and Royauté J. (2001). Utilisation d'un système d'analyse d'information dans le processus de la veille scientifique et technologique : acceptabilité et pratiques collaboratives induites. In *Actes du 3e Congrès du Chapitre français de l'International Society for Knowledge Organisation* (ISKO): 79-88.

Ibekwe-SanJuan F. and SanJuan E. (2003). TermWatch : cartographie de réseaux de termes. In *Proceedings of the 5th Conference on Terminologie et Intelligence Artificielle (TIA'03)*: 124-134.

Ibekwe-SanJuan F. and Dubois C. (2002). Can Syntactic variations highlight semantic links between domain topics? In *Proceedings of the 6th International Conference on Terminology and Knowledge engineering (TKE'02)*: 57-63.

Ibekwe-SanJuan F. (1998). A linguistic and mathematical method for mapping thematic trends from texts. In *Proceedings of the 13th European Conference on Artificial Intelligence (ECAI'98)*: 170-174.

Jacquemin C. (2001). *Spotting and discovering terms through Natural Language Processing*. MIT Press.

Lebart L. and Salem A. (1994). *Statistique textuelle*. Dunod.

Lelu A. (2001). Synthèse d'information en ligne : bilan du prototype NeuroWeb. In *Actes du 3e Congrès du Chapitre français de l'International Society for Knowledge Organisation* (ISKO): 187-195.

Reinert M. (1990). ALCESTE : une méthodologie d'analyse des données textuelles et une application : Aurélia de G. de Nerval. *Bulletin Méthodologie sociologique*, vol. (26).

SanJuan É. and Ibekwe-SanJuan F. (2002). Terminologie et classification automatique des textes. In *Actes des JADT 2002* : 677-688.

Small H. (1999). Visualizing science by citation mapping. *Journal of the American society for Information Science*, vol. (50/9): 799-813.

Silberztein M. (1993). INTEX© manual, *2000-2001*. ASSTRIL – LADL.

El análisis estadístico para el estudio de los campos estilísticos en una obra literaria

Angel Igelmo, Gabriel M. Jordà, Carlota Vicens

Universitat de les Illes Balears
Cra. de Valldemossa, Km 7,5 – 07071 – Palma de Mallorca – España
angel.igelmo@uib.es, gmjorda@uib.es, cvicens@uib.es

Abstract

The purpose of this page is to outline some of the possibilities offered by using the FRECONWIN tool in the textual analysis. Vocabulary can be seized as a dynamic whole ; we consider textuality as a network of thin mutual relations between recurrent units, a vocable is rather a field, a discontinous cotext, wich rules its own endogenous norm for the distribution of its coocurrents. The object is a statistical and stylistic investigation of 'Citadelle' by Antoine de Saint-Exupéry.

Resumen

El objetivo de esta comunicación es el de mostrar la fecundidad de los análisis estadísticos efectuados con la aplicación FRECONWIN en el estudio de los campos estilísticos de una obra literaria. Siguiendo a P. Guiraud, consideramos "campos estilísticos" los nudos de interrelaciones que se establecen entre los diferentes elementos del léxico de una obra. Hemos elegido la que Saint-Exupéry consideraba como la más importante de su vida, *Citadelle*, y con la finalidad de descubrir su estructura léxica y poder conocer mejor el pensamiento y el universo simbólico del autor hemos estudiado los campos estilísticos de los nombres, de los más específicos en alguna de las etapas de la vida del autor y también de los más homogéneos, de los que sistemáticamente aparecen a lo largo de la obra.

Keywords : statistique textuelle, homogénéité et spécificité, champs stylistiques, Saint-Exupéry.

1. Valores estilísticos del vocabulario

A menudo se ha utilizado la estadística léxica como rama auxiliar de una estilística de normativa exógena ; es decir, se tenía como representativo del estilo de un autor el conjunto de desviaciones en la frecuencia del empleo de determinados vocablos respecto a un corpus externo considerado como característico del uso de la lengua en un tiempo-espacio determinados (Guiraud, 1954). De este modo se esperaba de los métodos estadísticos, no sólo la identificación y comparación de las desviaciones, sino también su interpretación estilística, considerando como elemento esencial, como *mots-clés*, la lista de palabras "sobrerepresentadas". Sin embargo el mismo Guiraud insistió, unos años más tarde, en "la extrema complejidad del problema", constatando con sinceridad que "la mayor parte de los estudios ... de los *mots-clés* y de los *écarts* en el uso de las formas y construcciones son, por regla general, simples inventarios pasivos de los que sólo pueden extraerse conclusiones vanas o tautológicas" (Guiraud, 1969 : 16). Frente a este problema se produjo un cambio de perspectiva al considerar que los vocablos se caracterizan por su colocación a lo largo del texto (Peytard, 1970), por su "cotexto discontinuo" (Massonie, 1986 ; Muller, 1973 y 1977 ; Brunet, 1985) ya trabajaron desde esta perspectiva comparando cada obra de Corneille o cada novela de Zola con el conjunto del teatro dramático o

con el de los Rougon-Macquard ; por su parte Viprey estudia la dinámica del vocabulario en las relaciones de afinidad que se observan a través de su distribución a lo largo del texto, la "surreprésentation d'un vocable dans le cotexte d'un autre vocable du même ensemble" (Viprey, 1977 y 1998).

Ya dentro de esta perspectiva, nosotros hemos seguido el punto de vista de Guiraud cuando considera la lengua de una obra literaria como un sistema de valores en el cual los signos funcionan oponiéndose unos a otros, recibiendo un sentido a medida que se van forjando sus relaciones recíprocas dentro del conjunto de la obra. Surge así la noción de "campo estilístico" para definir los nudos de interrelaciones que se establecen entre los diferentes elementos del léxico de una obra. Nuestro trabajo se centra por consiguiente en el estudio de cada vocablo dentro del conjunto de la obra y, para ello, el análisis estadístico se inicia no sólo con la consideración de los elementos "sobrerepresentados", "específicos", sino también de los más "homogéneos", los que aparecen de forma sistemática a lo largo de toda la obra que, desde este momento, consideramos como un mensaje del que pretendemos reconstruir el código (Guiraud, 1969).

2. El corpus analizado

El objeto del análisis ha sido *Citadelle*, obra póstuma de Antoine de Saint-Exupéry. Apreciado en todo el mundo por su producción literaria y mal interpretado a veces por sus compatriotas, Saint-Exupéry es una figura literaria clave en el período de entreguerras. La aplicación utilizada, FRECONWIN, ha sido desarrollada en la Universidad de les Illes Balears con el objetivo fundamental de ofrecer los cálculos de homogeneidad y epecificidad de formas y segmentos léxicos, tanto en un corpus general como en una clase de vocablos de dicho corpus. La personalidad del escritor y la del ser humano están íntimamente ligadas ; coherente con su vida y con su obra, en una carta dirigida a André Breton que nunca llegó a enviar, afirma : "Il n'est pas une ligne écrite au cours de toute mon existence qu'il me soit nécessaire de justifier, de taire ou démentir" (1982 : 136). El manuscrito de *Citadelle* se encontró en el interior de un maletín que él confió a su compañero Gavoille la víspera de su muerte. *Citadelle* ha sido la obra más controvertida de Saint-Exupéry y la que ha provocado las valoraciones más polémicas, pero era la obra en la el autor que tenía depositadas todas sus esperanzas, "c'est l'œuvre de toute ma vie", afirmó al confiar el manuscrito a su amigo (Richelmy, 2000 : 129) ; fue comenzada hacia 1936, aunque algunos estudiosos sitúan el inicio de su redacción a 1933 y otros, asociándola a su experiencia del desierto en Cap Juby, a 1927. De lo que no cabe duda es de que fue escrita simultáneamente a *Terre des hommes*, *Pilote de guerre*, *Lettre à un otage* y *Le petit prince*, y que la parte substancial de ella se escribió durante su exilio en New York, es decir, cuando Europa se enfrentaba al totalitarismo, a la guerra y a la destrucción, no sólo física, sino también de los valores de la sociedad humana, por los que él siempre luchó. Según propia afirmación pensaba consagrar a la redacción del libro todavía diez años más para pulirlo y revisarlo. Estas características convierten el manuscrito de *Citadelle* en un corpus ideal para el estudio de su estructura léxica mediante el análisis estadístico y de los campos estilísticos ; Saint-Exupéry, contra lo que algunos piensan, era extremadamente minucioso en la corrección de sus obras, manuscritos de 400 páginas quedaban a menudo reducidas a unas 150 ; éste no es el caso de *Citadelle* ; esta obra póstuma es un compendio de notas manuscritas del autor o grabadas durante la noche para ser más tarde escritas, por lo que no puede en absoluto ser considerada desde el punto de vista formal como una obra acabada y, en consecuencia, sometida a comparación estilístico-formal con las otras obras del autor ; por el contrario, al ser un compendio de relatos y reflexiones surgidos en un primer momento de elaboración, efectuados a lo largo de casi toda su vida literaria y con los que Saint-Exupéry pensaba crear su "Biblia", según propias palabras, tienen para

nosotros el valor de la inmediatez y la espontaneidad, contituyen un testimonio inapreciable para, desde el análisis estadístico, poder descubrir la estructura léxica y penetrar en el pensamiento y universo simbólicos del autor ; el estudio de los campos estilísticos de los nombres más específicos y homogéneos nos permite conocer el proceso en que se van creando los dominios y subdominios conceptuales, un proceso del que Saint-Exupéry era muy consciente : "Et quand j'ai dit rempart il faut aussi remplir le mot. Et les géomètres y ajoutent quelque chose, et les poètes, et les conquérants, et l'enfant pâle et la mère [...]" (1971 : 107). La edición de *Citadelle* empleada es la publicada por Gallimard (1971), en su colección "Le Livre de Poche", nº 1.532, 1.533 y 1.534. El manuscrito, que su autor no pudo corregir, llegó a la editorial lleno de tachaduras, incorrecciones y líneas en gran parte ilegibles. Esta edición presenta el texto considerado hoy como definitivo gracias a la puesta a punto de Simone Lamblin, Pierre Chevrier y Léon Wencelius.

3. Análisis estadístico, campos estilísticos y estructura léxica

En la preparación del texto hemos escogido la opción de los "no lematizadores", "formalistes" (Lafon, 1984) en vez de la de los "lematizadores", "lemmatiseurs" (Muller, 1984) y, una vez realizada la desambiguación, dadas las características de las seiscientas páginas de *Citadelle*, un conjunto de pensamientos, reflexiones y notas que no fueron revisadas ni estructuradas por el autor, hemos escogido la "División automática" del FRECONWIN, optando por dividir el documento en cinco partes de tamaño idéntico. La elección de cinco partes es arbitraria en el sentido de que, dadas las características de *Citadelle*, no existe ninguna razón estilística, temática o biográfica para justificar una división determinada, la única razón es la estadística por lo que hemos optado por un número que sin ser demasiado bajo para un análisis estadístico aceptable, dividiera la obra en partes de una extensión significativa (39.621, 39.475, 39.476, 39.475 y 39.467 formas).

Una vez dividido el documento hemos iniciado nuestro análisis lexicométrico ; éste se ha centrado en el estudio de los "mots pleins", es decir, el nombre, el verbo, el adjetivo y el adverbio. Nuestro objetivo ha sido obtener las formas y segmentos más específicos y homogéneos para, a continuación y utilizando los medios que nos proporciona la aplicación FRECONWIN, encontrar sus campos estilísticos y elaborar la estructura léxica de la obra. Como se ha señalado, esta comunicación expone el estudio de los nombres.

3.1. *Análisis estadístico*

Para el estudio de la homogeneidad la aplicación sigue el método del Chi-2. Este contraste o comparación lo equiparamos a un contraste de homogeneidad de muestras. Se trata de contrastar la hipótesis nula :

$$H_0 : p_1 = p_2 = \ldots = p_k$$

con un nivel de significación α, frente a la hipótesis alternativa, H_1, de que alguna igualdad no se cumple, con nivel de significación α. El nivel de significación es la probabilidad de rechazar la hipótesis nula, H_0, siendo cierta. En este trabajo emplearemos $\alpha = 0'05$.

En este caso se trata de verificar si la clase gramatical analizada es homogénea en los k textos, lo cual significa, bajo la hipótesis de que H_0 sea cierta, que las diferencias entre las proporciones de los textos son debidas al azar y no representan un cambio de estilo.

Si $n_1, n_2, \ldots, n_k$ son las frecuencias absolutas de la clase gramatical analizada en cada texto y $N_1, N_2, \ldots, N_k$ son los tamaños de dichos textos, entonces el estadístico de contraste es :

$$W = \frac{1}{p(1-p)} \sum_{i=1}^{K} \frac{(n_i - pN_i)^2}{N_i}$$

siendo :

$$p = \frac{n_1 + n_2 + ... + n_k}{N_1 + N_2 + ... + N_k}$$

El estadístico de contraste tiene una distribución Chi-cuadrado con *k-1* grados de libertad :

$$W \sim \chi^2_{k-1}$$

lo cual permite determinar la región crítica, RC, o región de rechazo de tamaño α.

El criterio decisorio será por tanto : rechazar H_0 si $W \in RC$. El *nivel de significación* nos muestra la posibilidad de que se rechace la hipótesis nula aún siendo ésta cierta (según el estadístico, según la hipótesis desarrollada) ; dicho de otra forma, nos muestra el índice de error : una significancia de 0'05 quiere decir que tenemos una probabilidad de acierto de un 95%.

Las aproximaciones estadísticas para realizar el estudio de la especificidad han sido múltiples, basándose en teorías tales como *Chi-cuadrado*, *la ley normal* o *ley de Poisson*. Sin embargo, es la *ley hipergeométrica* aplicada al análisis de textos por Lafon (1980) la que se adapta mejor a las ocurrencias del vocabulario. La especificidad puede ser positiva, y por tanto haber una sobreutilización, o ser negativa lo que señalaría su escasez. Lafon desarrolló el cálculo de las especificidades siguiendo la ley de la distribución hipergeométrica y demostrando que ésta se adapta a la perfección al campo actual de la lexicometría. Éste es pues el método utilizado en el *FRECONWIN*.

Pretender atacar el problema mediante el cálculo directo de la distribución hipergeométrica (mediante números combinatorios), no es el método más adecuado para un ordenador, ya que los valores numéricos son demasiado elevados. El método que seguimos, aplica la escala logarítmica para intentar disminuir los números procesados en el ordenador.

Traduzcamos ahora el modelo mediante fórmulas matemáticas : convenimos en llamar a *T* a la longitud total del *corpus* y *t* a la longitud de una parte. El computo de las posibles muestras de longitud *t* es elemental : es el número de combinaciones que pueden ser formadas eligiendo *t* elementos entre *T*. Este número se representa por :

$$\binom{T}{t} = \frac{T(T-1).....(T-t+1)}{t(t-1).....2.1}$$, o expresado de otra forma : $$\frac{T!}{t!(T-t)!}$$

Sea ahora *f* la frecuencia total de una forma dada *F*. Del conjunto extraemos una muestra de longitud *t*. El número de ocurrencias de *F* en la muestra es una variable aleatoria *X* que puede tomar los valores *0,1,2, ..., k, ... f*. Podemos calcular la probabilidad para que *X=k*, es decir para que *F* figure exactamente *k* veces en la muestra.

Para esto calculamos las muestra que contiene *k* veces *F*. Provienen de todas las combinaciones posibles de *k* elementos entre *f*, es decir $\binom{f}{k}$, y para cada una de ellas de todas las combinaciones de *(t-k)* elementos entre *(T-f)* ; es decir :

$$\binom{T-f}{t-k}$$

En la hipótesis de la equiprobabilidad de las muestras la probabilidad buscada es pues :

$$\Pr ob(X = k) = \frac{\binom{f}{k}\binom{T-f}{t-k}}{\binom{T}{t}}$$

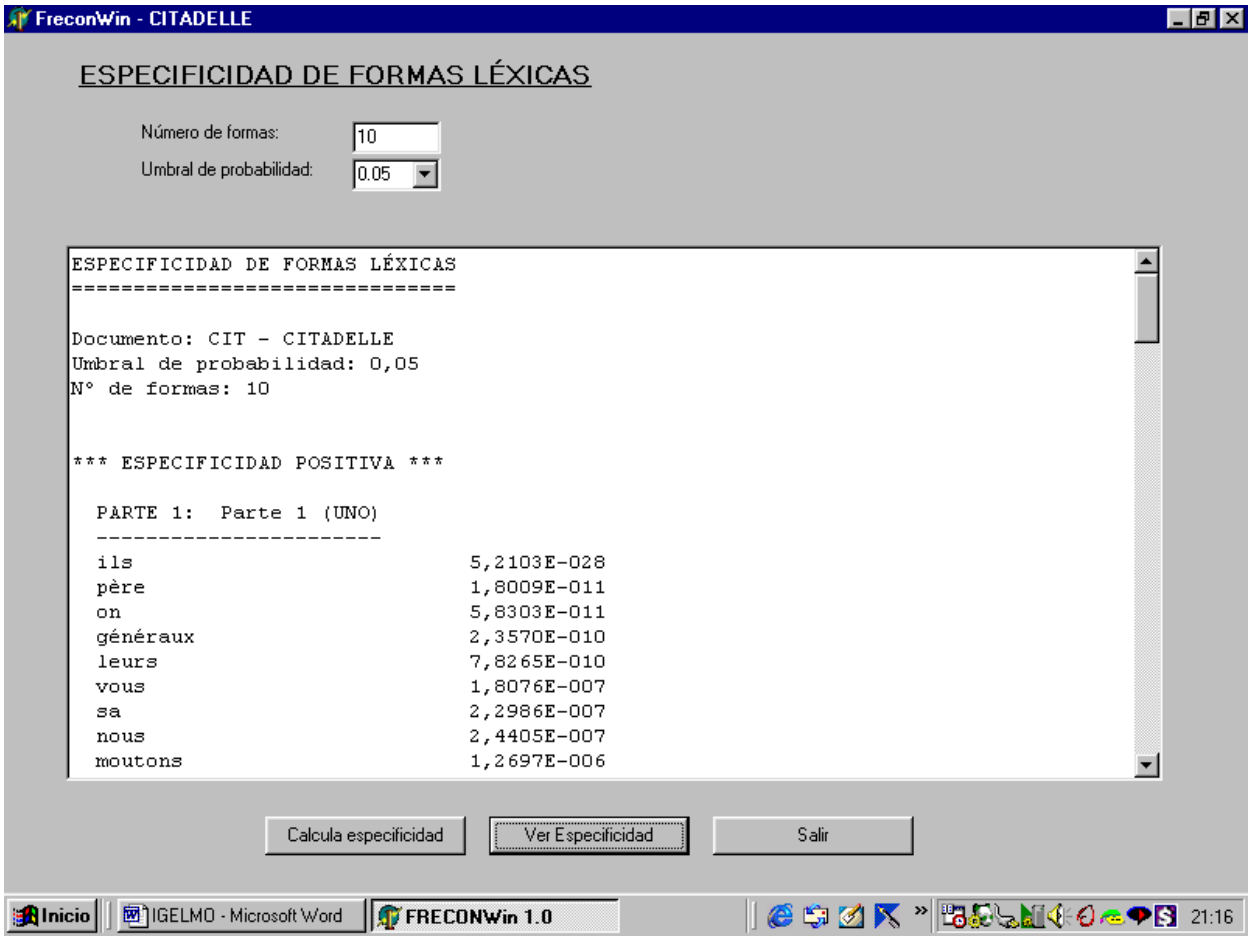
3.2. *Resultados del análisis estadístico*

3.2.1. *Especificidad*

La aplicación informática nos ha permitido la obtención de las formas específicas, sobreutilizadas e infrautilizadas, de cada una de las partes.

Ofrecemos a modo de ilustración una de las pantallas del cálculo de la especificidad con el FRECONWIN, aunque hay que indicar que éste nos permite también agrupar las formas por categorías, en el caso que nos ocupa, la de los nombres, y llevar a cabo los cálculos estadísticos dentro de cada una de ellas. En el análisis de los nombres en *Citadelle* respetamos la diferencia entre mayúsculas / minúsculas y singular / plural porque a menudo el autor utiliza estas marcas para otorgar distintos sentidos a las palabras, p. e. : *Dieu* / *dieu* / *dieux* ; *demeure* / *demeures* ; *Homme* / *homme* / *hommes* ; *Intelligence* ; *Esprit*.

A continuación ofrecemos los resultados de los diez nombres sobreutilizados e infrautilizados en cada una de las cinco partes en las que hemos dividido la obra de Saint-Exupéry.



3.2.1.1. *Los nombres de 'Citadelle' : especificidad positiva y negativa*

Parte 1 (páginas 15-133)

Positiva :

père ; *généraux* ; *demeure* ; *moutons* ; *demeures* ; *temple* ; *or* ; *sédentaires* ; *collaboration* ; *sel.*

Negativa :

Cérémonial ; *clous* ; *poème* ; *ami* ; *lignes* ; *planches* ; *sentinelle* ; *roi* ; *heure* ; *étage*.

Parte 2 (páginas 134-253) :

Positiva :

vie ; *paysage* ; *crête* ; *loisir* ; *prière* ; *plan* ; *création* ; *ordre* ; *Dieu* ; *montagne*.

Negativa :

choses ; *cérémonial* ; *condition* ; *sentinelle* ; *remparts* ; *blé* ; *graine* ; *part* ; *besoin* ; *matériaux*

Parte 3 (páginas 254-273) :

Positiva :

sens ; *liberté* ; *sentinelle* ; *choses* ; *temple* ; *empire* ; *âmes* ; *réalité* ; *contrainte* ; *objets*.

Negativa :

Soir ; *instant* ; *signe* ; *tour* ; *condition* ; *mot* ; *façon* ; *morts* ; *part* ; *bonheur*.

Parte 4 : (páginas 374-496) :

El análisis estadístico para el estudio de los campos estilísticos en una obra literaria

Positiva :

Graine ; *part* ; *ElKsour* ; *champ* ; *roi* ; *géant* ; *qualité* ; *change* ; *sucs* ; *homme*.

Negativa :

Vie ; *cause* ; *sourire* ; *femmes* ; *besoin* ; *demeure* ; *corps* ; *or* ; *frère* ; *pouvoir*.

Parte 5 : (páginas 497-617) :

Positiva :

Perle ; *condition* ; *frère* ; *fête* ; *cérémonial* ; *Seigneur* ; *soif* ; *amour* ; *bijou* ; *échecs*.

Negativa :

Sens ; *langage* ; *homme* ; *liberté* ; *hommes* ; *forme* ; *porte* ; *Dieu* ; *armée* ; *part*.

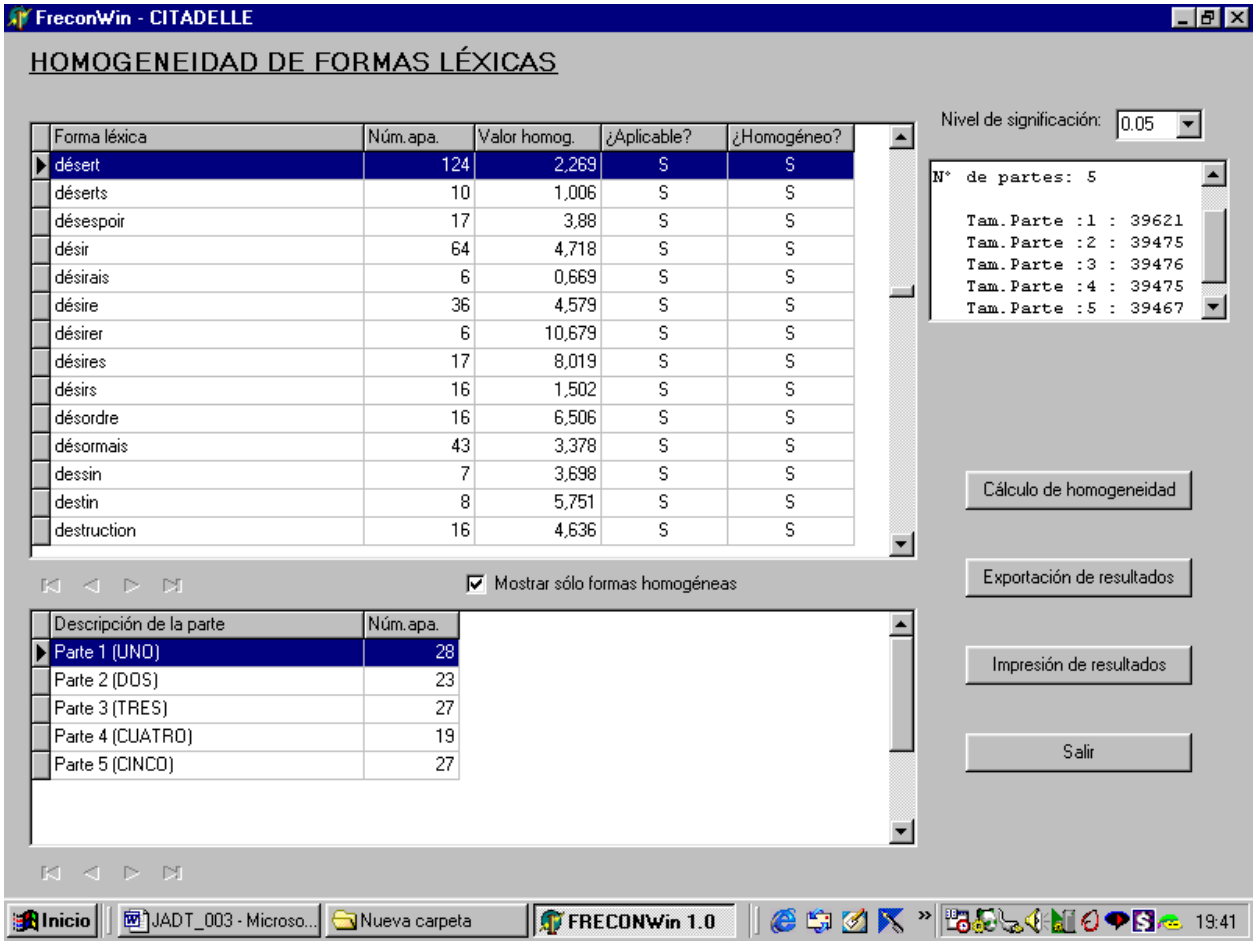
Ya de entrada es posible hacer las siguientes observaciones :

- Las formas más utilizadas para cada una de las partes son las siguientes : *père*, *vie*, *sens*, *graine* y *perle*, frente a las menos utilizadas : *cérémonial*, *choses*, *soir*, *vie* y *sens*.
- La forma *vie* es la más utilizada en la segunda parte y la menos utilizada en la cuarta. Lo mismo ocurre con *sens*, la más utilizada de la tercera parte y la menos de la quinta.
- Las formas sobreutilizadas ofrecen en general un índice de especificidad más homogéneo.
- Ninguno de los nombres sobreutilizados se repite en ninguna de las partes. De entre los nombres infrautilizados se repiten los siguientes : *cérémonial* en la 1ª y 2ª partes, *sentinelle* en la 1ª y 2ª partes, *condition* en la 2ª y 3ª, *besoin* en la 2ª y 4ª y "part" en la 2ª, 3ª y 5ª.
- Al agrupar los nombres según campos semánticos se observa que, a grandes rasgos, estos coinciden tanto para los nombres más utilizados como para los menos utilizados. Las refe-

rencias a la religión, con palabras como *Dieu, cérémonial, temple, prière, âme* o *Seigneur* (con mayor presencia en el grupo de las formas sobreutilizadas), así como las referencias al *habiter*, habitar la casa (*demeure, père, frère, amour, loisir, ami* ...) o habitar la *citadelle* (*sentinelle, empire, roi, remparts, tour, porte* ...) son lás más frecuentes, aunque los nombre empleados no siempre son los mismos. Por otra parte en el grupo de las formas sobreutilizadas observamos abundantes referencias a la naturaleza (*paysage, montagne, graine, moutons, champ* ...), mientras que en el grupo de las infrautilizadas llaman la atención aquellas formas que se refieren a la comunicación (*lignes*, *poèmes, mot, signes, langage, sens* ...).

3.2.2. *Homogeneidad*

A continuación hemos obtenido el índice de homogeneidad de los nombres con un umbral de frecuencia cien o superior, con un nivel de significancia de 0'05 y 4 grados de libertad ; el resultado ha sido que hemos rechazado la hipótesis nula en 18 de ellos. Aunque como cabía esperar, ninguna de las formas homogéneas coincide con ninguna de las sobre o infrautilizadas, es interesante constatar que los campos semánticos de unos y otros siguen siendo muy próximos. Así, seis de los 18 nombres se refieren a elementos de la naturaleza y un segundo grupo a la ciudadela que hay que habitar. Por el contrario no se observan palabras directamente relacionadas con la religión y ninguna referencia al mundo de la aviación.



Forma léxica	Núm.apa.	Valor homog.	¿Aplicable?	¿Homogéneo?
désert	124	2,269	S	S
déserts	10	1,006	S	S
désespoir	17	3,88	S	S
désir	64	4,718	S	S
désirais	6	0,669	S	S
désire	36	4,579	S	S
désirer	6	10,679	S	S
désires	17	8,019	S	S
désirs	16	1,502	S	S
désordre	16	6,506	S	S
désormais	43	3,378	S	S
dessin	7	3,698	S	S
destin	8	5,751	S	S
destruction	16	4,636	S	S

Descripción de la parte	Núm.apa.
Parte 1 (UNO)	28
Parte 2 (DOS)	23
Parte 3 (TRES)	27
Parte 4 (CUATRO)	19
Parte 5 (CINCO)	27

Los nombres de uso más homogéneo en 'Citadelle'

Forma léxica	Núm.apa.	Valor homog.	¿Aplicable ?	¿Homogéneo?
fois	175	001.71	S	S
désert	124	001.78	S	S
vent	124	002.23	S	S
maison	164	003.39	S	S
peuple	113	004.62	S	S
marche	107	004.79	S	S
monde	122	004.88	S	S
arbre	262	005.32	S	S
visage	196	005.50	S	S
enfant	132	006.04	S	S
jour	170	006.29	S	S
vérité	134	006.42	S	S
chose	186	007.14	S	S
mots	126	007.29	S	S
travail	106	007.49	S	S
mer	171	008.01	S	S
pierre	108	009.07	S	S
yeux	102	009.10	S	S

3.3. *Estudio de los campos estilísticos en 'Citadelle'*

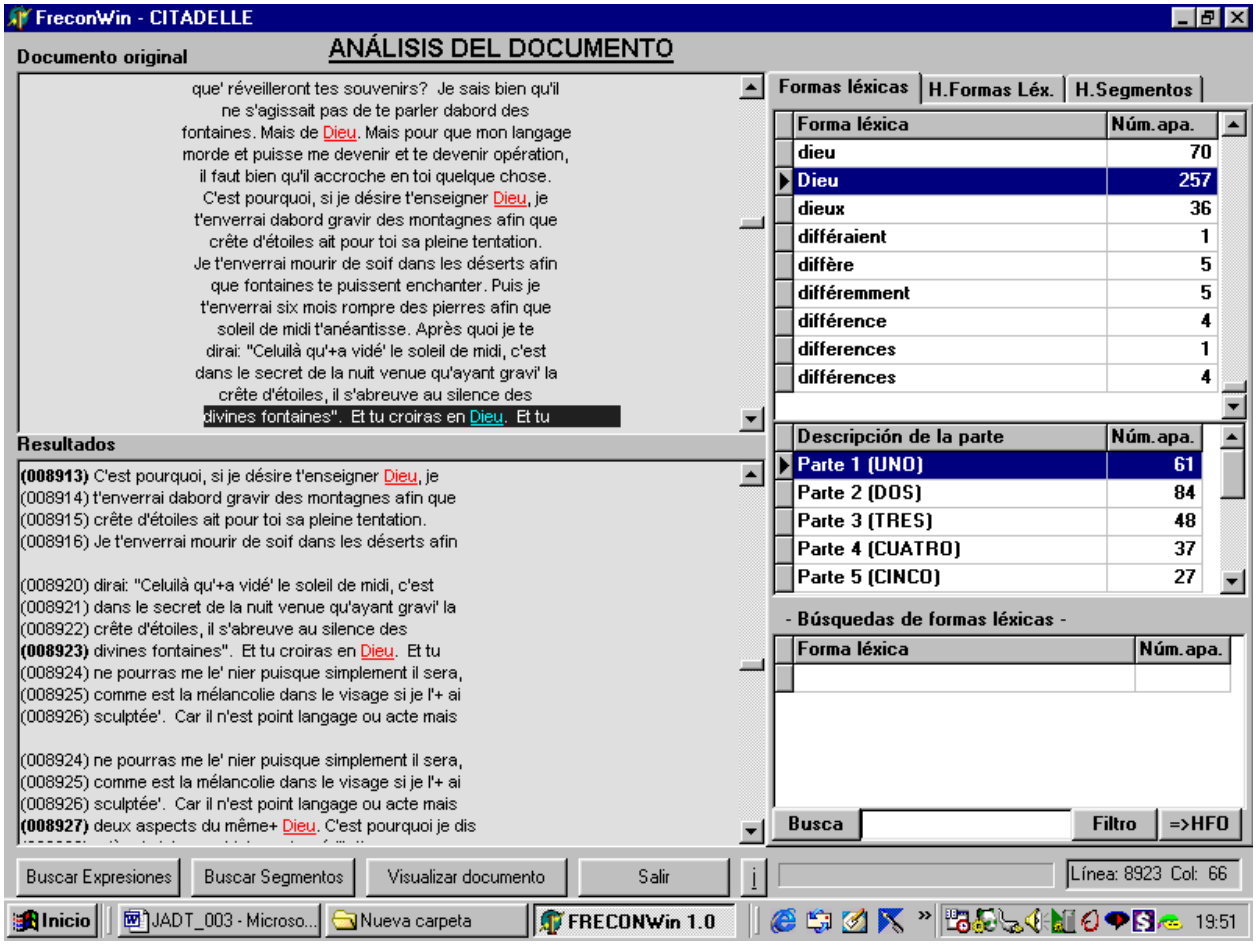
Finalmente hemos decidido adentrarnos en el estudio de los campos estilísticos y de la estructura léxica de *Citadelle* analizando los segmentos (específicos y homogéneos) junto con los respectivos contextos de 33 nombres, los tres más sobreutilizados de cada una de las cinco partes y de los dieciocho homogéneos : *père, généraux, demeure, vie, paysage, crête, sens, liberté, sentinelle, graine, part, El Ksour, perle, condition, frère, fois, désert, vent, maison, peuple, marche, monde, arbre, visage, enfant, jour, vérité, chose, mots, travail, mer, pierre, yeux.*

La aplicación FRECONWIN nos permite agrupar los segmentos que se articulan en torno a una "forma polo", descubrir cuáles son los que aparecen de una forma más recurrente y cuáles sólo lo hacen en un contexto léxico determinado, de este modo se articula la red que constituye el campo estilístico de la "forma polo" inicial y del que va brotando el significado de la palabra dentro de la estructura de *Citadelle*.

Del estudio de los campos estilísticos se deriva la absoluta interdependencia por una parte de los nombres de la obra en su construcción del centro ; por otra de los diferentes símbolos en su caminar hacia este símbolo último que gobierna el paisaje citadeliano. Hacia el centro concurren efectivamente símbolos ascensionales, catamorfos, espectaculares, diurnos o nocturnos, cíclicos, diaréticos ..., que pueden reducirse a dos grupos según su relación con el mismo : los que se alejan de él, fuerzas centrífugas o de desintegración (*frère*, *généraux*, *sentinelle*, *vent*), y los que conducen a él, fuerzas centrípetas o de integración (*arbre*, *graine*, *désert*, *mer*, *perle*, *paysage*, *crête*, *yeux*, *visage*, *pierre*, *maison*, *demeure*, *peuple*, *père*). La llegada a este centro es inevitable para quien sigue los caminos de las ascensión, el *échange* y el *habiter*.

De esta forma el estudio lexicométrico de *Citadelle* nos ha permitido descubrir una estructura léxica en la que se manifiesta un universo simbólico que guarda gran coherencia con el resto de

la obra exuperiana, que se configura como una obra unitaria y circular. Así la imagen del *Espíritu* (con mayúscula) que impera en *Citadelle* como lo único capaz de construir al Hombre, se encuentra ya intuída o bocetada en obras anteriores (*Courrier sud*, *Terre des hommes*), lo mismo que la diferencia entre *Cuerpo/Inteligencia/Espíritu* (*Pilote de guerre*) o (*Petit prince*) el viaje entendido como camino iniciático. Es cierto que sorprende en *Citadelle* la ausencia del avión y del concepto de vuelo, pero el autor no renuncia sin embargo a la mirada desde un lugar elevado, al paisaje contemplado desde la altura, finalmente organizado y comprendido.



Referencias

Guerrero J.L., Igelmo A. y Jordà G.M. (2003). *Freconwin*. Palma de Mallorca. Universitat de les Illes Balears.

Guiraud P. (1954). *Caractères statistiques du vocabulaire*. P.U.F.

Guiraud P. (1969). *Essais de Stylistique*. Klincksieck.

Massonie J.-P. (1986). "Q-occurrences libres". In Brunet Ét., *Méthodes quantitatives et informatiques dans L'étude des* textes. Slatkine-Champion : 611-623.

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Slatkine-Champion.

Muller Ch. ([1977] 1992a). *Principes et méthodes de statistique lexicale*. Champion.

Muller Ch. ([1977] 1992b). *Initiation aux méthodes de la statistique linguistique*. Champion.

Muller Ch. (1984). Prologue. In Lafon P., *Dépouillements et statistiques en lexicométrie*. Slatkine-Champion.

Peytard J. y Genouvrier É. (1970). *Linguistique et enseignement du français*. Larousse.

Richelmy M. (2000). *Homme synthèse du siècle*. Éditions lyonnaises d'art et d'histoire.

de Saint-Exupéry A. (1971). *Citadelle*. Gallimard. Folio.

de Saint-Exupéry A. (1982). *Écrits de guerre 1939-1944*. Gallimard. Folio.

Sánchez Hernández Mª A. (2001). *El verbo en 'Citadelle' (A. de Saint-Exupéry). Análisis estadístico, campos estilísticos y estructura léxica*. Tesis de doctorado. Las Palmas de Gran Canaria : Universidad de Las Palmas de Gran Canaria.

Vicens Pujol C. (2000). *Los nombres de 'Citadelle' de Antoine de Saint-Exupéry : análisis estadístico, campos estilísticos y estructura léxica*. Tesis de doctorado. Barcelona : Universitat de Barcelona, División de ciencias humanas y sociales, Departamento de Filología Románica.

Viprey J.-M. (1997). *Dynamique du vocabulaire des 'Fleurs du Mal'*. Slatkine-Champion.

Viprey J.-M. (1998). "Une norme endogène pour le calcul stylistique du vocabulaire". In *Actes des JADT 1998.*