TyPTex : Generic Features for Text Profiler

G. Illouz, B. Habert, H. Folch, S. Heiden, S. Fleury, P. Lafon, S. Prévost
LIMSI B.P. 133 F−91403 Orsay cedex, France
& UMR 8503 – ENS de Fontenay/Saint–Cloud,
2 avenue de la Grille du Parc F−92211 Saint−Cloud, France
{habert,illouz}@limsi.fr,
{fleury,folch,heiden,lafon,prevost}@ens−fcl.fr

Abstract

Very large corpora are increasingly exploited to improve Natural Language Processing (NLP) Systems. This however implies that the lexical, morpho−syntactic and syntactic homogeneity of the data used are mastered. This control in turn requires the development of tools aimed at text calibration or *profiling*. We are implementing such profiling tools and developing an associated methodology within the ELRA benchmark named *Contribution to the construction of corpora of contemporary French*. The first results of this approach – applied to a sample of the main sections of *Le Monde* newspaper – yields constraints for corpus profiling architectures.

Résumé

Le recours croissant aux "très grands corpus" pour améliorer les systèmes de Traitement Automatique des Langues (TAL) suppose de maîtriser l'homogénéité lexicale, morpho−syntaxique et syntaxique des données utilisées. Cela implique en amont le développement d'outils de calibrage de textes. Nous mettons en place de tels outils et la méthodologie associée dans le cadre de l'appel d'offres ELRA *Contribution à la réalisation de corpus du français contemporain*. Nous montrons sur les rubriques principales du journal *Le Monde* les premiers résultats de cette approche. Nous précisons les contraintes qui en résultent pour les chaînes de traitement de corpus, au regard des propositions existant dans le domaine.

1 Introduction: Corpus profiling.

1.1 Change of paradigm in NLP

Nowadays, a significant improvement of NLP systems is expected mainly through the exploitation of very large corpora (Amstrong, 1994), via the acquisition of linguistic knowledge which needs to be on the one hand very vast (vast, in terms of the number of lexical entries and rules) and on the other hand very detailed (concerning words' syntactic usage constraints or their privileged associations, for instance).

The availability of lexical, syntactic and semantic textual data in NLP has gained huge proportions (100 million tagged words of the BNC – *British National Corpus* [1], for instance), but this data is at times extremely heterogeneous. This is the case for corpora whose source are press articles, such as the CD−ROM of *Le Monde* newspaper, which, given its availability, is often used for developing NLP tools for French. It is nevertheless necessary to profile the texts that are available in order to check the correspondence between the language uses that are represented in them and the ones that the NLP tools being developed are intended to tackle.

1.2 Mastering the data used for acquisition

Different studies converge to show that the reliability of NLP treatments in different domains depends on the homogeneity of the data which is being used.

1 This corpus http://info.ox.ac.uk/bnc/ includes spoken (10%) as well as written language (fiction from 1960 onwards and "informative" texts from 1975 onwards). It fulfills the aim of providing a set of textual data whose production and reception conditions are precisely defined and which is representative of a great variety of communication situations. Cf. (Habert *et al.*, 1997, p. 147)(Kennedy, 1998, p. 48–50).

Tagging D. Biber used corpora divided into different domains (in the LOB – *Lancaster−Oslo−Bergen*) [2] to show (Biber, 1993, p. 223) that the probability of occurrence of a morpho−syntactic category is a function of the domain. D. Biber also showed (*ibid.*, p. 225) how the sequence of probabilities of morpho−syntactic categories varies across domains. Similarly, collocation was shown to differ from one domain to another (for instance for *sure* and *certain*). When aiming at developing a probabilistic tagger, limiting one's scope to a single domain introduces a heavy bias to the training process. Simply summing up results over different domains leads to inexploitable averages.

The precision of the taggers which have been evaluated within the framework of GRACE (Adda *et al.*, 1999), measured in relation to the manually tagged reference corpus, similarly shows significative variations depending on the part of the corpus which is under examination (Illouz, 1999). This corpus containing 100,000 words has been compiled from extracts from *Le Monde* (2 extracts), and from literary texts: memoirs (2 extracts), novels (6 extracts), essays (2 extracts). Thus an extract from memoirs results in important variations, positive and negative, among the taggers.

Parsing Sekine (Sekine, 1997) uses 8 domains of the Brown corpus (documentaries, editorials, hobbies, learned, fiction, western, romance novels). He examines the performances, measured in recall and precision, of a syntactic probabilistic parser when the training is performed on the same domain as the one used for the test, when it is performed on all domains, when it is performed on the fiction part (fiction, western, romance novels) or on the non−fiction part (documentaries, editorials, hobbies, learned)(he calls the two latter groupings *classes*). The different performances are in decreasing order: training corpus domain equals test corpus domain, training corpus domains and test corpus domains belong to the same class, training corpus and test corpus are composed of extracts from all domains. Training the parser on a class (fiction, for instance) and using it on the other class (non−fiction) yields the worst results.

Information retrieval Kalgren (Karlgren, 1999, p. 159–161) uses the part of the TIPSTER corpus [3] corresponding to the *Wall Street Journal* and the queries 202 to 300 of the TREC (*Text Retrieval Conference*) evaluation campaign together with the relevance judgements concerning the 74,516 articles in question [4], in other words, the indication of whether the article is a correct response or not to the given query. It measures a certain number of stylistic features of each article: average word length, proportion of long words, average word frequency, average frequency of capitalised words, proportion of digit letters, personal pronouns, etc. It emerges that the texts which are judged relevant differ significantly from the texts judged not relevant, and further that the texts most frequently selected by systems in competition within TREC (either relevant or not) also differ significantly from the texts for which no relevance judgement exists.

As is revealed by these experiences covering the different fields of tagging, parsing, and information retrieval [5], a corpus can induce two kinds of statistical errors which threaten the validity of any generalisation that is derived from it (Biber, 1993, p. 219–220): random error and bias error. *Random error* occurs when a sample is too small to represent the global population. *Bias error* occurs when one or several characteristics of a sample are systematically different from those of the population that the given sample is supposed to represent [6]. In the wake of these kinds of errors it appears necessary to

2 This tagged corpus was designed to be the english counterpart to Brown. Brown is a tagged corpus of a million words which has been made available in 1979 by W. Francis and H. Kucera at the Brown University (USA). The Brown corpus is composed of 500 extracts consisting of 2,000 occurrences each. These extracts have been compiled from american texts published in 1961 corrresponding to 15 different "genres": journalism, scientific and technical texts, etc. It has been carefully tagged. Due to its public availability, it has played an important part in the renewed interest in corpus studies. LOB is equally a 1 million word corpus which has been constructed according to the same criteria, the only difference being that it has been compiled from english texts published in 1961.

3 http://www.tipster.org

4 These articles date from 1990 to 1992. 2,039 are relevant for at least one query. 35,289 are not relevant for one single query. 37,188 articles remain unjudged.

5 The experiment in (Pichon & Sebillot, 1999) shows that corpus homogeneity could be as well at stake in Word Sense Disambiguation.

6 Y.Wilks and R. Gaizauskas (Wilks & Gaizauskas, 1999, p. 198) have commented on the *de facto* privilege bestowed on the *Wall Street Journal* among recent work in information retrieval: "the over use of one single type of text for training may have had profound but unmeasureable effects as of yet on this field: the linguistic

become well−acquainted with the strictly linguistic characteristics of the corpus (Pery−woodley, 1995).

2 Profiling methodology and architecture

2.1 Definition

We call *text profiling* the use of tools aimed at "calibrating" a corpus in terms of its vocabulary as well as in terms of categories and morpho−syntactic patterns. The aim of calibration is to produce measures of homogeneity within the different parts of a corpus in terms of one or more parameters.

Our approach consists in developing a typology of texts through inductive methods. In other words, our text types are defined in terms of sets of correlated lexical and syntactic which have been extracted through multi−variate statistical techniques from annotated corpora. This approach is based on D. Biber's work (Biber, 1988)(Biber, 1995). Biber uses 67 features corresponding to 16 different categories (verb tense and aspect markers, interrogatives, passives, etc.). He examines their distribution in the first 1,000 words of 4,814 contemporary english texts. The identification of the 67 features in the corpus is done automatically on the basis of a preliminary morpho−syntactic tagging. The sets of correlated features (the dimensions) are obtained through a multi−variate statistical technique (factor analysis). Each dimension consists of 2 complementary groups of features which can be interpreted as positive and negative poles. When one group of features occurs in a text, the other group is far less frequent. Any new text can then be placed in the space of *n*−dimensions previously identified. The location of the new text along each dimension in this space is determined by the frequency count of each of the features of the dimension. Clustering methods are then being used to group texts in terms of their location in this space. The resulting clusters are types of texts which correspond directly neither to text "genres" nor to language styles or registers.

Contrastive studies on the performance of NLP tools on such text types will make it possible to test their robustness in relation to the linguistic variation among the different types and to determine for which type each tool is specifically adapted . One can then adjust the treatments in consequence.
Inversely, calibration tools will make it possible to position a new text in relation to the text clusters which have already been found and thus help choose the most adequate treatments.

We develop this work within the TyPTex project financed by ELRA (*European Language Resources Association*), and we carry it out jointly at LIMSI and at UMR 8503. It consists of developing a methodology and a toolkit aimed at testing and extending Biber's work, based on the results obtained for French by Sueur (Sueur, 1982) and Bronckart (Bronckart *et al.*, 1985). The 3 projects which have been selected for this benchmark will use a common corpus which has 5 million words, of which 1 million come from *Le Monde* newspaper and constitute a subset of the *Press* section of the PAROLE corpus (see below).

2.2 Global architecture of the TyPTex project

As shown in figure 1, at the bottom level, the architecture consists of a collection of texts which are tagged according to the TEI (*Text Encoding Initiative*) recommendations. Each text has a descriptive header attached to it (Dunlop, 1995). We then perform queries to extract a subset of texts which are relevant to a certain study or application. The next step is to perform a morpho−syntactic tagging which associates each lexical item (or poly−lexical item) to a given word stem, a part of speech category and other morpho−syntactic information. We then perform *typological marking*. It consists of replacing the information generated by the morpho−syntactic tagger by higher−level categories. These new categories are calculated from the morpho−syntactic tags and vary according to which features we want to study. From the resulting tagged corpus (possibly corrected by CorTecs (Heiden *et al.*, 1998)) several matrices are generated, in particular the matrix containing the frequencies of each feature in each text of the corpus under study. The resulting matrix is then analysed by statistical

treatment of the *Wall Street Journal* has undoubtedly improved, but it is not certain that it can unlock the secrets of all the world's texts!"

software programs. The analysis of the matrix is aimed, on the one hand, at identifying the relevant features of a certain opposition, and on the other hand, at making an inductive or supervised classification of texts.

We use currently Sylex-Base (Sylex, 1995) for tagging. It is a tagger/parser based on the work of P. Constant (Constant, 1991) which proved to be robust during the tagger evaluation programme GRACE. The typology tagging, which is at the present day still limited, includes shifters, modals, presentatives, tense use, passives, certain classes of adverbs (negation, degree), determiners, etc. The category (or part of speech) is kept for those words or polylexical items which have not been otherwise tagged.

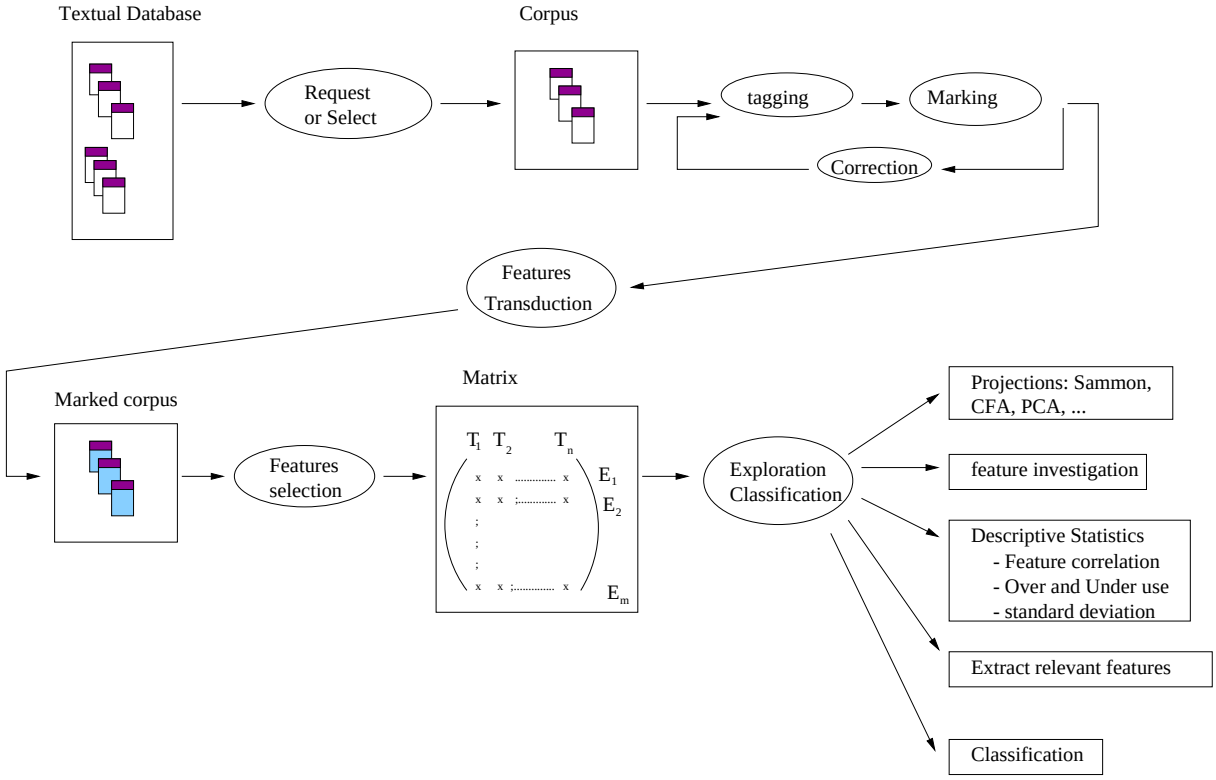


Figure 1: Text profiler Architecture

3 Testbed: getting acquainted with *Le Monde*

The current series of tests are limited to *Le Monde* papers, as the rest of the corpus [7] for the ELRA was not available for the conference deadline. However these experiments cover the different functionalities of the present TyPTex architecture. They mainly help in defining the necessary improvements.

3.1 Data under study and previous experiences

The corpus which has been put together by G. Vignaux (InaLF – *Institut National de la Langue Française*) and B. Habert within the framework of the european project PAROLE comprises one sub−part *Press* of 14 million words which has been constructed by random selection of whole issues of *Le Monde* newspaper. It gathers issues dating from 1987, 1989, 1991, 1993 and 1995 (Naulleau, 1998).

7 European Parliament debates, books and papers in social sciences and arts, scientific popularization, novels.

A pilot study (Illouz *et al.*, 1999) was conducted on the following 6 sections – the most important in terms of size: ART [8] (arts, media, entertainment), ECO (economy), EMS (?education, health, society), ETR (foreign affairs), ING (?general information, sports, current affairs), POL (politics). It showed that there were significant differences across these sections, both concerning the employed vocabulary as well as the favoured syntactic categories [9].

The current study concerns the same sections. But only articles comprising 1,000 to 2,000 words have been taken into account so as to avoid comparing articles whose size is too dissimilar [10]. The sub−corpus under study has a total number of occurrences of 2,160,071. Each one of the articles has been tagged by `Sylex-Base`. The resulting tags have been substituted with typological categories (see above).

3.2 Distinctive sections but similar articles

We have used the computation of *characteristic elements* (Lafon, 1980)(Lebart *et al.*, 1998, p. 130–136) to find the significant over and under use of a category in a given section in relation to its distribution in all 6 sections and to the length of the whole corpus. In table 1, a rectangle on the right hand side of an axis indicates an over−use, and a rectangle on the left hand side, an under use, the size of the rectangle is in inverse proportion to the probability of this over or under use (the greater the rectangle, the more characteristic the element).

From the approximately 200 features marked at present, we have kept about 40, divided into 2 subsets. The first subset comprises the functional elements whose role is to organise the discourse and the sentence, the second subset comprises open categories: nouns, adjectives, verb tenses...

The grammatical apparatus in ART manifests an expressive, even an affective textual structuring: over−use of exclamation marks, question marks, and of presentatives (present indicative of *c''est* and *il y a*), graded adverbs, as well as pronouns, in particular shifters (1st person singular and 2nd person plural). ECO is almost at the opposite end. POL favours the use of colon, of the subordinate *que*, as well as negative adverbs: indicators of reported speech (direct and indirect) and of controversial discourse ?

The adjectives oppose ART, ECO, EMS and ETR (with an over−use) to ING, and POL which avoid them. The significant over−use of personal nouns in POL is probably favoured by the presence of election results. Nominalisations (*the trapping of X* versus *X traps Y*) – hence the same kind of agent deletion – are common to both ECO and POL. A certain distance towards what is being reported – by way of an over use of *Pouvoir* (*can* / *may*) present conditional – brings together ECO and ETR, while POL reveals obligation [11] (*Il Faut* and *Devoir* in present indicative) and ART displays both over and under uses for these modal forms. Verb tense also varies across sections. ART favours the opposition *Présent* (present) / *Passé simple* (simple past) (and under−use of *passé Composé*). ECO favours participles, future and present conditional (is there a guarded announcement of some news?); EMS uses future tense intensively; ETR over employs the past tense, as is also the case for ING which moreover is lacking in future and present tense occurrences; POL displays a significant lack of any tense except the present. Passive opposes the over uses of ECO and EMS to the under uses of ART and POL, whilst ETR and ING display both over and under uses.

---

8 The descriptive fields of the `PAROLE` corpus adopt the classification codes used by the *Le Monde* newspaper. We had not access to the meaning of all these codes, hence the question marks.

9 In as far as the vocabulary is concerned, 15,438 articles comprising a total of 7 million words pertaining to the relevant sections were under study. As for the syntactic categories, 241,484 words belonging to 7 issues dating back to september 1987, were examined. They were extracted from the previous set, tagged automatically and the part of speech corrected manually.

10 The articles in the super−set corpus comprise from 13 to 5,202 words, 455 in average.

11 Checking the contexts would be necessary to see whether it is an interpretation of eventuality of *Devoir* which appears in ECO and one of necessity in POL.

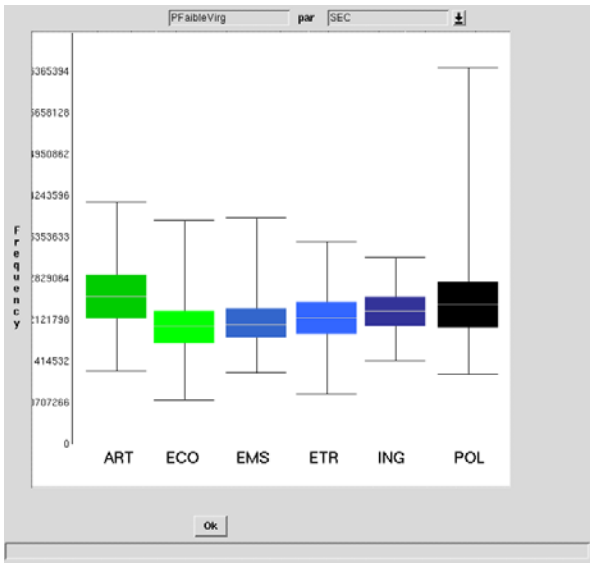
Feature	ART	ECO	EMS	ETR	ING	POL
Commas						
Colons						
Quotes						
Exclamation marks						
Interrogation marks						
Periods						
Conjunctions « que »						
Conjunctions (“relatifs”)						
$C'est_{present\ indicative}$						
$Il\ y\ a_{present\ indicative}$						
Coordinations						
Prepositions						
Adverbs						
Degree adverbs (« très »,. . .)						
Negation adverbs						
Pers. pron. 1st pers. sing.						
Pers. pron. 2nd pers. sing.						
Pers. pron. 1st pers. plur.						
Pers. pron. 2nd pers. plur.						
Pers. pron. 3rd pers.						
Other pronouns						
Definite determiners						
Indefinite determiners						
Prep. + definite det. (« des »,. . .)						
Adjectivess						
Cardinal determiners						
$Nouns_{common\ nominalisation-}$						
$Nouns_{common\ nominalisation+}$						
$Nouns_{proper}$						
$Devoir_{present\ indicative}$						
$Il\ faut_{present\ indicative}$						
$Pouvoir_{present\ conditional}$						
Future indicative						
Imperfect indicative						
Present indicative						
Present conditional						
“Passé composé“						
“Passé simple“						
Past Participle						
Present Participle						
Passive future indicative						
Passive infinitive						
passive “PasséComposé”						
Passive indicative present						

Table 1:Stylistic oppositions among the 6 main sections of *Le Monde*

This clear distinction between the different sections taken as a whole is opposed to the global stylistic homogeneity of the articles when taken one by one. Indeed, when each article is represented by a vector composed of the number of occurrences of the 200 features selected for stylistic purposes, the contrasts between sections are clearly less marked, as is shown in figure 4. We have therefore developed tools aimed at observing with greater accuracy the features according to the classification schemes used by *Le Monde* newspaper's editorial board. The first one of these being the possibility to make the total feature occurrences relative to one given feature. Hereafter, this operation has been carried out for word counts.

3.3 Distribution of feature frequencies

Let's look at the features *comma* and *proper nouns* for instance, marked as distinctive by the specificity analysis for nearly each section. Their distributions are shown in figures 2a and 2b. These figures display the lower extreme, the lower quartile, the median, the upper quartile and the upper extreme, which are represented by the five horizontal marks.



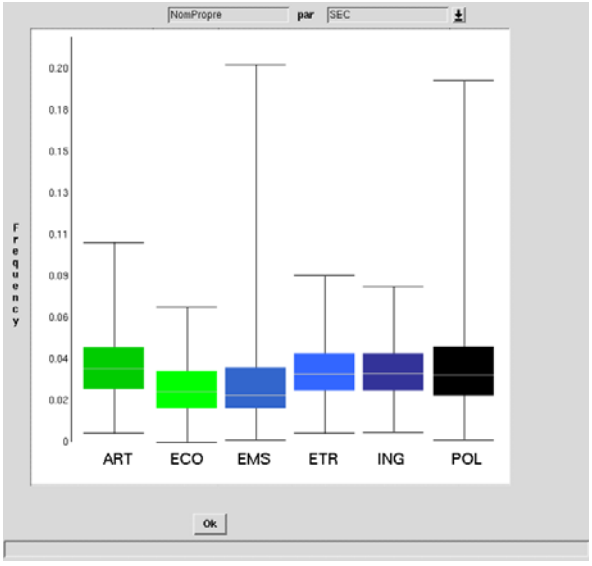


Figure 2: (a) Comma Distributions (b) Proper Noun Distributions

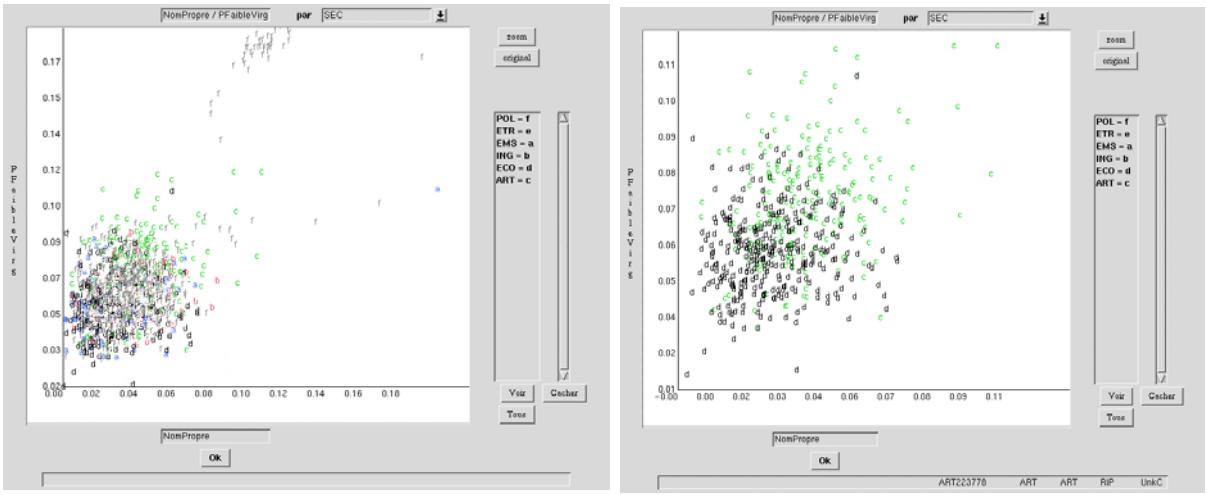
One can see that although the features' distribution is different among the different sections, the range of values for the features are, on the whole, similar; or in any case it exists for each text class displaying identical or similar values for a given feature. For a given feature to be relevant to a given classification on its own, it would need to partition the range of values for each class. We have checked this for each feature.

After an examination of each one of the 200 features, we have not obtained any partition in terms of classes. We have then explored the correlation of features taken two at a time, though not exhaustively (as there are about 2,000 possible combinations).

3.4 Exploration of correlations features tuples

Let's turn again to the two features of the previous example and examine their correlation, in other words study the cloud of dots (articles), by classes (characters *a*, *b*, *c*, *d*, *e*, *f*). Figure 3a shows the results for all sections. In the top centre is a cluster of POL texts (character *f*), corresponding to election results. Given that the rest of figure 3a is difficult to interpret, we also present the results for the classes ART and ECO (figure 3b). This latter chart clearly shows that the centres of gravity of the two classes are different, thus confirming the specificity analysis. However the border remains fuzzy

and therefore precludes inducing a classification.



(a) All sections (b) ART and ECO

Figure 3: Feature Comma *versus* Feature Proper Noun

Several tuples have thus been examined (for instance noun / adjective to check whether adjectivisation is proportionally constant). However no strongly discriminant tuples have emerged. We have therefore used multi−dimensional techniques; at present, we have used the *Sammon projection* (Sammon, 1969)[12].

3.5 Multi−dimensional analysis of feature distributions

The feature vectors associated to each article are explored using Sammon's method, which projects a *n*−dimensional cloud of points onto a *k*−dimensional space ($k < n$), while preserving the distances existing in the space of origin in the best way possible. In figure 4, the points which correspond to articles do not appear grouped in distinct regions in accordance to the section they belong to. In (Illouz *et al.*, 1999), the same projection method was used on samples of sections comprising 10,000 words, and features defined by the 108 lexical forms whose count was superior to 500 occurrences. In that case, clusters of articles by sections clearly emerged. Maybe the observation window (1,000 to 2,000 words) here is too fine−grained to produce observable phenomena.

This contradiction between the global differences across sections and the groupings of articles taken individually does not however indicate the existence of text types defined as correlations of features, in the corpus under study. It is necessary to turn to multidimensional statistics to pursue this hypothesis.

This experience of typological marking has enabled us however to examine the features we have used in a critical light. They can be too fine−grained and lead to a scattering of occurrences which makes contrasts imperceptible. This has been the case concerning verb tenses in the current choice of features: the verb category is fragmented into some 50 features, most of which have a limited number of occurrences. Therefore we have no grip on the verb considered globally, nor on its tendencies with respect to the sections or to the articles. We have no way of knowing, for example, whether the under−use of nouns in POL is related to an over−use of verbs, as had been observed for a subset of the same corpus in (Illouz *et al.*, 1999), which would be in accordance with the under use of all verb tenses shown in table 1. Inversely, certain features are too rough and probably hide real oppositions. This is the case for cardinal numbers that groups quantity indicators, as well as dates, which would probably be more effective to differentiate. We would in fact wish to manipulate feature structures in

12 Correspondance analysis is under development at the moment, to be integrated in the analysis module of the `TyPTex` architecture

order to be able to use the corresponding information totally or partially [13]. Thus for instance, having {category=noun,type=common,nominalization=no,gender=masculin,person=singulier...} kind of tag enables us to select subsets such as {category=noun}, {category=noun, type=common}, or {gender=masculin}. Using feature structures such as those employed in unification grammars (Shieber, 1986) would make it possible to modelise more precisely the information resulting from marking, in the style of (Gazdar *et al.*, 1988) for instance, as well as the operations that can be performed on them [14].

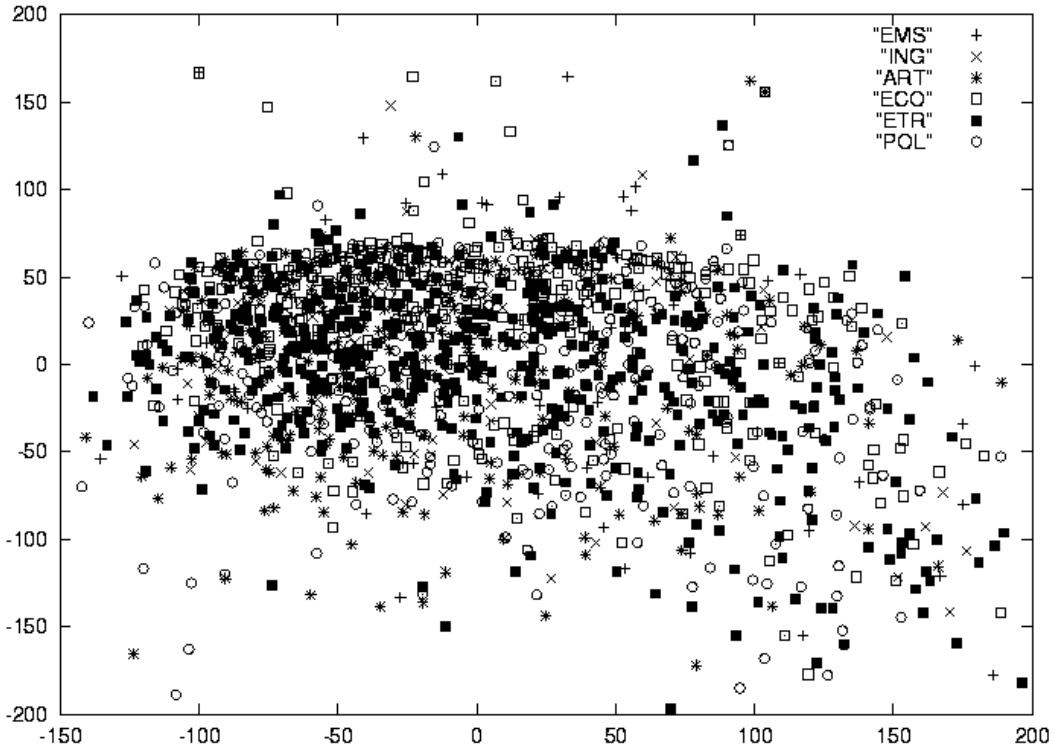


Figure 4: All stylistic features for 6 main sections articles

Following this approach, we have tested the possibility to sum features up, thus creating so called « super features ». Thus, we have constructed from elementary features, the following super−features: *Verb*, *Adverb*, *Substantive*, *Personal pronoun* and *Article* [15]. Next we have analysed them by a Sammon projection of the classes. As the classes were still not significantly distinctive, we have studied the classification by "genre". Although the latter is less present in *Le Monde* [16], it is however more subtle in view of observing certain stylistic distinctions and tuning our profiler. The corresponding Sammon projection is shown in figure 5 where only the most relevant classes with a sufficient number of occurrences have been kept. We can observe the emergence of a "market" grouping of texts that occupies a different space from obituary column or"chronology texts. The 2 latter have a comparable stylistic behaviour.

13 This is the approach of (Habert & Salem, 1995).

14 They are not constrained enough at present. It is not possible to assess the soundness of the marking.

15 Ressorting to feature structures, as presented in the previous paragraph, would permit to change a marking tag into the POS subsuming it. However, summing up features is still needed. For instance, it can be necessary to gather tags corresponding to the same function but resorting to different morpho−syntactic categories, such as some punctuation marks and some conjunctions as textual markers.

16 Only 15% of articles are attributed to a genre.

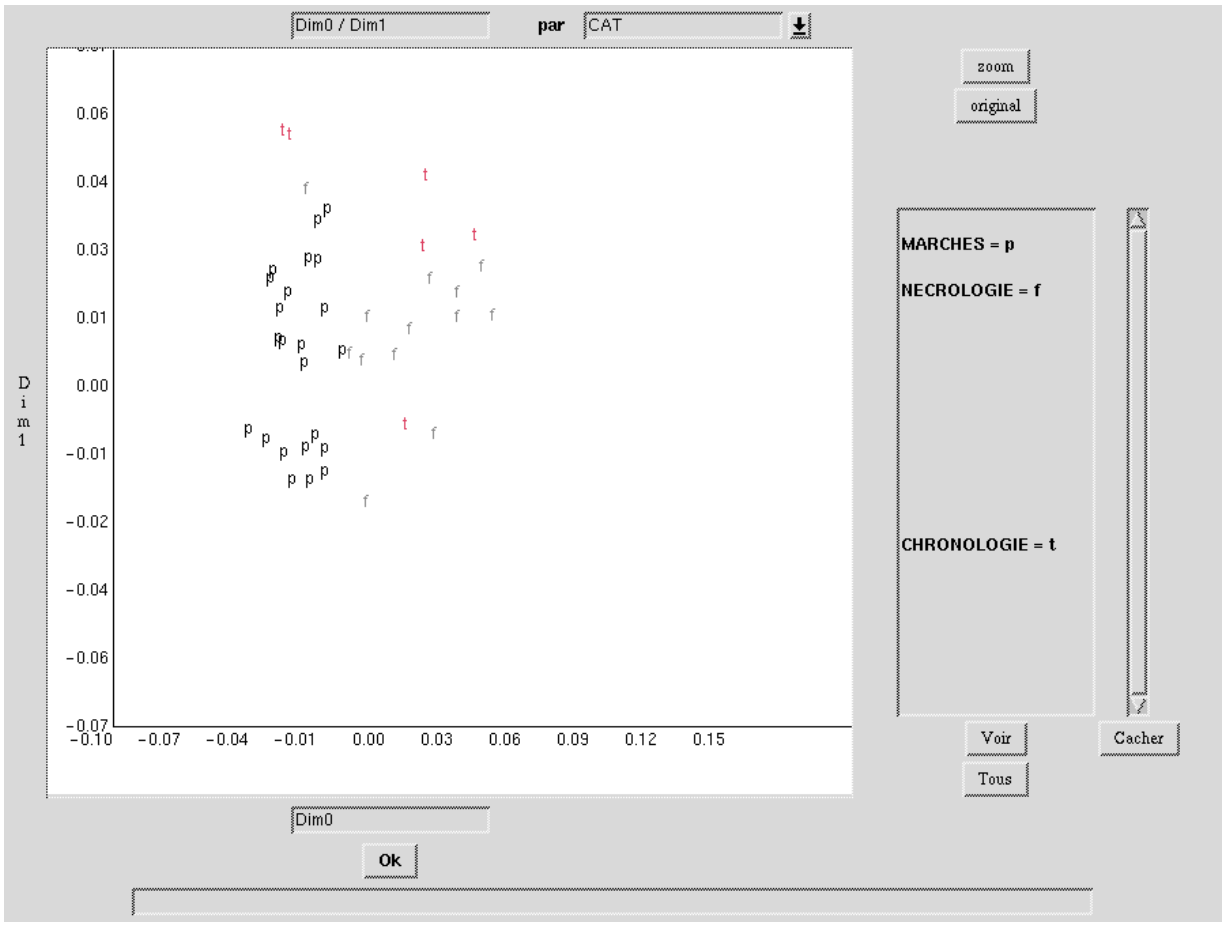


Figure 5: Sammon Projection on *text genre*

It is necessary therefore to be able to group features for one contrast, to divide others, at times even to start afresh tagging and marking for certain points. Corpus profiling requires flexibility in the manipulation of the corpus which in turn introduces constraints on the software architectures in use.

4. Evaluation of architectures for corpus processing

Paradoxically, it has been the exchange of language data – primary or tagged – that has concentrated most of the standardisation initiatives within the `TEI` framework (Ide & Veronis, 1995). In contrast, several paths have been examined for corpus processing toolkits. Two main problems can be distinguished (McKelvie *et al.*, 1997): how to deal with very large corpora accompanied with complex and sometimes contradictory [17] annotations (morpho−syntactic or semantic tagging, parse trees, co−reference markers...); and how to articulate software components in a modular way? In section 4.1 we present 3 architectures that we have tested and describe the way they tackle these problems and in section 4.2 we present our own choices which adapted to our particular tasks.

4.1 3 architectures: GATE, IMS−CWB, LT NSL

We have tested the following 3 architectures [18]:

17 Two morpho−syntactic taggers run on the same corpus can result in two different segmentations of lexical forms et diverge moreover on the analysis of certain segments.

18 At LIMSI, G. Illouz has used `GATE` to test the consequences of substitution a module (like tagging) of one software program by another. At UMR 8503, S. Heiden has integrated the query engine `CQP` of `IMS−CWB` and the `SGML` architecture `LT NSL` as advanced query tools for the program he has developped called `Lexploreur` (Heiden, 1999) for the exploration of textual data.

GATE (Wilks & Gaizauskas, 1999) This architecture is based on the need to make heterogeneous NLP modules intercommunicate for the development of complex systems. Annotations are stored separately from the primary data to which they refer. The graphical interface makes the task of interconnecting components and exploring different combinations of existing modules easier. However an effort is required to develop tools to generate an intermediate format [19] from the specific formats accepted and generated by existing modules.

IMS Corpus Workbench (Christ, 1994) This workbench is based on a search engine specifically designed for tagged corpora. Textual data, whose words or idioms can be tagged with any information, is first compiled and stored in a database comprising various indexes and symbol tables : that is why the IMS workbench is very effective for search in corpora whose tagging is stabilised. Then, it is possible to search for occurrences of complex textual events in the corpus with the CQP search engine (*Corpus Query Processor*). A CQP search expression lets one first express various constraints on the properties of words with Boolean combinations of regular expressions stating conditions on tagging information (like the POS – *Part of Speech* – of words, their lemma, their form...). Those single word occurrence expressions can then be combined in regular expressions on the occurrences of the first level of expression. This second level lets one express various patterns of co−occurrence of words in sentences, paragraphs [20]. The precision of a search expression depends on the way the corpus is tagged and on the correctness of the tagging.

It is interesting to note that the base result of any CQP search expression is a list of sequences of contiguous word positions in the corpus. This allows to exploit and display the search results in various ways: as hierarchical indexes (a list of word sequences sorted by sequence frequency), as KWIC concordances, distribution graphs...[21].

LT NSL / XML (McKelvie *et al.*, 1997) LT XML is basically an implementation of an XML parser and search engine. The search engine permits to go to and fro very smoothly between an event model of the XML stream (start then end of an element, #PCDATA occurrence...) and an XML element tree model (occurrence, at one node, of an element having certain values for some attributes, this node dominating another element following other constraints). The LT XML package includes various XML document processing tools, based on the LT XML library API [22]. The two main advantages of that architecture are: the possibility to express the processing of textual data at any level of detail of information encoding in the corpus (the properties of a single word through its attribute−value pairs for example, the title or number of a section, bibliographic information found in the header of a text...); the formal correctness of the XML/SGML data stream through all the processing.

Comparison of the three architectures Two solutions are thus available for the use of multiple annotations: storage of the annotations in a single document (IMS−CWB) *versus* distribution of the annotations (GATE). The first approach facilitates the subsequent access to the documents and the establishment of connections between the different levels of annotation. The second one is

19 Each atomic annotation is composed of a type word, group, phrase..., attributes, and their starting and ending positions in the original data.

20 For example, the CQP expression
[lemma="have"]([POS="adverb"][POS="adverb"][POS="subst.*pronoun"])[tense="preterit.*"] looks for some preterit constructions in the corpus : “any conjugation of the verb *have* followed by two successive adverbs and by a preterit verb”, or “the verb *have* followed by a substantive or a pronoun and followed by a preterit verb”. Square brackets contain the first level of single word occurrence expressions. Round brackets express grouping and the pipe sign expresses disjunction. The .* characters terminating some regular expressions on word properties express “any ending” of the property value.

21 Like in the integration of the CQP search engine in the Lexploreur software (Heiden, 1999) :
http://diderot.lexico.ens-fcl.fr/doc/lexploreur/index.html

22 The previous release of this software (called LT NSG) could validate and compile any SGML document in a *Normalised SGML format*. That format is based on an explicit SGML model: all opening and closing tags are explicit, all entities are expanded...All the LT NSG tools only worked on data in that format. The main features of this NSG normalised format are: it optimises parsing and search; it permits to combine successively the processing of different *SGML* tools, like in a Unix pipe, the data stream always being correct normalised SGML (see the integration of LT XML in the LML software (Heiden, 1999)). Since the advent of XML, the XML language format naturally replaced the NSG format in the library. Processing is now combined through XML data streams (validated or not).

favoured when the annotations diverge. It enables the articulation of a great number of simultaneous annotations. Furthermore, linking components one after another can be done using a pivotal format between 2 modules (GATE) – each module remaining "in control of itself" – or by rendering each module to a single format. The first solution favours the joint use of heterogeneous modules, the second one the homogeneity of the treatments.

4.2 Constraints associated with profiling

We specify for each profiling stage, the constraints to take into account and their compatibility with the three architectures presented above.

Varying the corpus From a single set of texts, multiple distinct corpora can be constituted. With this aim H. Folch has developed [23] a tool for corpus construction which builds a corpus from a query that articulates constraints on descriptive variables defined in each document header. In order to obtain stable text types, it is necessary to construct multiple corpora to examine the stability of the groupings resulting from the profiling tools [24]. GATE and IMS−CWB fix the state of the collection of texts under study once and for all and therefore do not fully satisfy this requirement.

Evolving features The features used in text profiling evolve. Firstly, the feature set is not closed. Several levels of representation are involved:

- Characters: punctuation marks, capital letters and digits in particular(Illouz, 1999);
- Closed lexical sets: categories of functional words(Brunet, 1981) (Biber, 1988)(Illouz *et al.*, 1999);
- Finer−grained typological categories(Sueur, 1982)(Bronckart *et al.*, 1985) (Biber, 1988);
- Text structure, title organisation, image presence, charts (Karlgren, 1999).

Secondly, at the bottom level, the morpho−syntactic tagging tools involved may render the identification of one or other of these features more or less difficult, and even in some cases lead to giving up on some of them (for instance for agentless passives). One can also resort to several of them.

Finally, what we still need to empirically determine are the set of features that are effectively discriminant and which can lead to clearly defined text types. This instability precludes the use of IMS−CWB during the tuning phase.

Multiple statistical treatments There are two kinds. The first type are aimed at exploring the significant correlations of linguistic features (Principal Component Analysis, Correspondence Analysis, Sammon). They consist of observing one feature or a small group of features in order to determine their relevance in relation to a classification. It enables the observation of features which are not necessarily ruled by the same probability laws (Karlgren, 1999, p. 153). This implies being able to visualise texts as points in a space, being able to change the point of view, the classification. The second type is that of supervised training. It implies being able to place a text in a pre−existent classification (via Quinlan's C4.5, for instance).

The tools which have been described here are to be found among different communities (data analysis, automatic learning) and are therefore difficult to use simultaneously. GATE can, in principle, articulate them, but encapsulates them to guarantee an interoperability of the treatments employed. To simplify this problem we have made the choice of defining a single contingency matrix where the rows are the texts and the variables (columns) are the features. We can thus provide each tool with the

23 Within the framework of the Scriptorium project of internal social watch carried out at the R&D division of EDF (*Électricité de France*). This project is developped through a joint contract valid during 1997−2000 signed between the R&D division of EDF and ENS Fontenay/St Cloud. It consists of the construction of a 20 million word corpus. The documents composing the corpus are extremely varied both in relation to their format as well as to the linguistic data they display: pamphlets, book extracts, corporate press, union press, summaries of corporate meetings, transcriptions of taped trade union messages, etc.

24 As carried out by Biber in (Biber, 1995) where he compares the results obtained for 4 different languages.

required data [25].

A permanent link between the primary texts and the profiling results Correlations between features resulting from multi−dimensional treatments are often hard to interpret (Karlgren, 1999, p. 157), as the example in table 1 reveals. It is essential, in order to check the proposed interpretations, to be able to examine the behaviour of these features within the context of the texts gathered together for the construction of the corpus. Moreover, certain results obtained through the profiling tools can be re−injected into the descriptive headers of the texts of origin. None of the architectures described above, is well adapted to implement this kind of backward link.

5 From a profiling prototype to a generic architecture

We are now in possession of a corpus profiler prototype. Globally, it supports the following treatments explicited in figure 1.

Corpus constitution A search engine exploits the descriptive fields of the text base to construct reasoned corpora.

Typological marking Transformation of the results of a morpho−syntactic tagger to focus on the linguistic behaviour that can underlie text types.

Exploration of feature tuples Visualisation of documents in terms of the marks constituted in this way. These can reveal clear contrasts between types of texts.

Articulate linguistic/metatextual descriptive information in order to examine the correlation among the different clusters obtained on the basis of linguistic features and of pre−existing classification schemes [26].

Combine features in order to obtain a "coarser" grain.

Pending tasks at this point are:

Using other taggers such as Tree Tagger [27] or Cordial 6 universités.

Improving the typological features in use Current and previous tests show the blind spots of those used at present (Illouz *et al.*, 1999).

Software integration The different phases of processing are not yet articulated into a coherent architecture like `LT NSL`.

Feature structuring It must enable the use of a "variable geometry" ranging from very coarse features (parts of speech) to very fine−grained tags (*Il faut* present indicative) or cut−across groupings (genre and person, for instance).

Bringing to light feature correlations Exploratory techniques will be put into place with this aim in mind.

A new project, `TyPWeb`, in collaboration with `CNET` (*Centre National d'Étude des Télécommunications*), aims at adapting the `TyPTex` architecture to the processing of web sites and will mark the passage of the present prototype to a generic profiling architecture. The aim of this project is to provide a methodological and practical framework for web site profiling and the development of a fine−grained typology of these sites. The approach consists of characterising each site by a set of indicators concerning both content and structure. The first step is to define and subsequently enrich the description of sites in terms of content and structure indicators: this

25 By extracting sub matrices, for instance.
26 Like the "genres" defined by *Le Monde* documentation: interview, obituary column, etc.
27 `http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html`

information is pumped into the descriptive header of the analysed sites. The header remains open and extendable by any new information deemed relevant. `TypWeb` should subsequently lead to a proposition of a content typology (using predefined topic indexes or constructing new content categories by way of an inductive approach). The resulting analysis should be obtained by crossing the formal structure with the content typologies. It will also consist of describing the articulation between the formal and semantic description of the sites with the practical account of the agents involved (designers and visitors). This approach aims in particular at analysing the progressive establishment of implicit exchange rules over the web (Beaudouin & Velkovska, 1999).

References

Adda, G., Mariani, J., Paroubek, P. & Lecomte, J. (1999). Métrique et premiers résultats de l'évaluation GRACE des étiqueteurs morphosyntaxiques pour le français. In P. Amsili, Ed., *Actes de TALN'99 (Traitement Automatique des Langues Naturelles)*, pp. 15–24, Cargèse: ATALA.

S. Amstrong, Ed. (1994). *Using Large Corpora*. Cambridge, Massachusetts: The MIT Press.

Beaudouin, V. & Velkovska, J. (1999). Constitution d'un espace de communication sur Internet (forums, pages personnelles, courrier électronique...). *Réseaux*, **17**(97), 121–177.

Biber, D. (1988). *Variation accross speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1993). Using register−diversified corpora for general language studies. *Computational Linguistics*, **19**(2), 243–258.

Biber, D. (1995). *Dimensions of register variation : a cross−linguistic comparison*. Cambridge: Cambridge University Press.

Bronckart, J.−P., Bain, D., Schneuwly, B., Davaud, C. & Pasquier, A. (1985). *Le fonctionnement des discours : un modèle psychologique et une méthode d'analyse*. Lausanne: Delachaux & Niestlé.

Brunet, E. (1981). *Le vocabulaire français de 1789 à nos jours, d'après les données du Trésor de la langue française*, volume I of *Travaux de linguistique quantitative*. Genève/Paris: Slatkine/Champion. Préface de Paul Imbs.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX'94 (3rd Conference on Computational Lexicography and Text Research)*, Budapest, Hungary. CMP−LG archive id 9408005.

Constant, P. (1991). *Analyse syntaxique par couches*. Doctorat de l'ENST, École Nationale Supérieure des Télécommunications, Paris.

Dunlop, D. (1995). Practical considerations in the use of TEI headers in large corpora. *Computers and the Humanities*, (29), 85–98. Text Encoding Initiative. Background and Context, edited by Nancy Ide and Jean Véronis.

Gazdar, G., Pullum, G. K., Carpenter, R., Klein, E., Hukari, T. E. & Levine, R. D. (1988). Category structures. *Computational Linguistics*, **14**(1), 1–19.

Habert, B., Nazarenko, A. & Salem, A. (1997). *Les linguistiques de corpus*. U Linguistique. Paris: Armand Colin/Masson.

Habert, B. & Salem, A. (1995). L'utilisation de catégorisations multiples pour l'analyse quantitative de données textuelles. *TAL*, **36**(1–2), 249–276. Traitements probabilistes et corpus, Benoît Habert (resp.).

Heiden, S. (1999a). Encodage uniforme et normalisé de corpus. Application à l'étude d'un débat parlementaire. *Mots*, (60), 113–132. Presses de Sciences Po.

Heiden, S. (1999b). *Manuel utilisateur du Lexploreur – Version 2.3*. Technical report, UMR 8503 – Analyse de corpus, Saint−Cloud.

Heiden, S., Cuq, A., Ducout, D., Horlaville, P., Robert, J.−P., Prieur, V. & Dohm, B. (1998).

*CorTeCs – 1.0*b *: Manuel de l'utilisateur*. Laboratoire de Lexicométrie et Textes Politiques – UMR 9952, CNRS – ENS Fontenay/Saint−Cloud.

N. Ide & J. Véronis, Eds. (1995). *The Text Encoding Initiative: Background and context*. Dordrecht: Kluwer Academic Publishers.

Illouz, G. (1999). Méta−étiqueteur adaptatif : vers une utilisation pragmatique des ressources linguistiques. In P. Amsili, Ed., *Actes de TALN'99 (Traitement Automatique des Langues Naturelles)*, pp. 15–24, Cargèse: ATALA.

Illouz, G., Habert, B., Fleury, S., Folch, H., Heiden, S. & Lafon, P. (1999). Maîtriser les déluges de données hétérogènes. In A. Condamines, C. Fabre & M.−P. Péry−Woodley, Eds., *Corpus et traitement automatique des langues : pour une réflexion méthodologique*, pp. 37–46, Cargèse.

Ingenia (1995). *Manuel de développement Sylex−Base*. Ingenia – Langage naturel, Paris. 1.5.D.

Karlgren, J. (1999). Stylistic experiments in information retrieval. In T. Strzalkowski, Ed., *Natural Language Information Retrieval*, pp. 147–166. Pays−Bas: Kluwer.

Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *MOTS*, (1), 128–165. Presses de la Fondation Nationale des Sciences Politiques.

Lebart, L., Salem, A. & Berry, L. (1998). *Exploring Textual Data*. Text, Speech and Language Technology. Dordrecht: Kluwer Academic Publishers.

McKelvie, D., Brew, C. & Thompson, H. (1997). Using SGML as a basis for data−intensive NLP. In *Proceedings 5th Conference on Applied NLP*, pp. 229–236: ACL.

Naulleau, E. (1998). *Tranformation of* Le Monde *data to obtain PAROLE DTD conformance*. Technical report, INaLF – CNRS, Saint−Cloud.

Pichon, R. & Sébillot, P. (1999). Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience. In P. Amsili, Ed., *Actes TALN'99*, pp. 279–288, Cargèse: ATALA.

Péry−Woodley, M.−P. (1995). Quels corpus pour quels traitements automatiques ? *TAL*, **36**(1–2), 213–232. Traitements probabilistes et corpus, Benoît Habert (resp.).

Sammon, J. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computing*, (18), 401–409.

Sekine, S. (1998). The domain dependence of parsing. In *Fifth Conference on Applied Natural Language Processing*, pp. 96–102, Washington: Association for Computational Linguistics.

Shieber, S. N. (1986). *An Introduction to Unification−Based Approaches to Grammar*. CSLI Lecture Notes 4. Stanford, CA: CSLI.

Sueur, J.−P. (1982). Pour une grammaire du discours : élaboration d'une méthode; exemples d'application. *MOTS*, (5), 145–185.

Wilks, Y. & Gaizauskas, R. (1999). Lasie jumps the GATE. In T. Strzalkowski, Ed., *Natural language information retrieval*, Text, speech and language technology, chapter 8, pp. 197–214. Dordrecht: Kluwer.