Building a database of French frozen adverbial phrases

Aude Grezka and Céline Poudat
LDI UMR 7187, CNRS / University of Paris 13, France
University of Paris 13, France
aude.grezka@ldi.univ-paris13.fr, poudat@univ-paris13.fr

Abstract

The present paper gives an account of the approach we have led so far to build a database of frozen units. Although it has long been absent from linguistic studies and grammatical tradition, linguistic frozenness is currently a major research issue for linguistic studies, as frozen markers ensure the economy of the language system.
The objective of our study is twofold: we first aim to build a comprehensive database of completely frozen units for the French language – what is traditionally called absolute or total frozenness. We started the project with the description of adverbial units – in the long term, we will also naturally describe adjectival, verbal and nominal phrases – and we will first present the database we have developed so far.
This first objective is necessarily followed by the second one, which aims to assess the frozenness degree of the other units (*i.e.* relative frozenness). In this perspective, we resorted to two sets of methods: linguistic tests and statistical methods processed on two corpora (political and scientific discourse).

Keywords: lexical database, frozenness, adverbials.

1. Introduction

Although it has long been absent from linguistic studies and grammatical tradition, linguistic frozenness is currently a major research issue for linguistic studies. Recent investigations (François & Mejri, 2006; Mejri, 1997; Grossmann & Tutin, 2003; Svensson, 2004) have shown that frozenness is at stake in language description, as frozen markers ensure the economy of the language system. Frozenness is indeed a massive phenomenon that should be considered as a definitional property of natural languages. For these reasons, all linguistic theories encompass the notion, which is both a descriptive and methodological concern. Frozen constructions vary a great deal internally, and the markers are numerous and need to be identified. This would moreover be particularly useful for NLP systems, which lack accurate methods to identify these markers.

The objective of our study is twofold: we first aim to build a comprehensive database of completely frozen units for the French language – what is traditionally called absolute or total frozenness. Absolute frozenness is nevertheless an exception: it would concern one unit out of ten on average, as G. Gross (1996: 16) pointed it out "variants are more common than total frozenness". This first objective is necessarily followed by the second one, which aims to assess the frozenness degree of the other units (*i.e.* relative frozenness). In this perspective, we will resort to two sets of methods: linguistic tests and statistical methods processed on corpora. Relative frozenness includes any frozen unit that accepts at least one variation within the free combinatorial system. Yet the definition has to be refined, as frozenness cannot be considered without the notion of degree, which depends on a set of semantic and syntactic criteria.

The building of a large database including all types of frozen units (adverbial, adjectival, nominal and verbal) will allow us to observe frozenness using a set of syntactic and semantic criteria ranked by relevance: this will enable us to assess the frozenness degree of the units, as well as the variant possibilities (in the case of relative frozenness). This will enable users and NLP systems to identify frozen units, and linguists to understand the mechanisms leading to total frozenness.

The present communication will present the approach we have led so far to build a database of frozen adverbial units. After a presentation of the database (Section 2), we will expose and discuss directions of ongoing work on the assessment of a frozenness degree using linguistic tests (Section 3) and statistical measures (Section 4). This section will also propose research directions to extend and evaluate the database coverage using corpora.

2. Presentation of the database

2.1 Building of the database

The FixISS database contains about 5000 frozen or semi-frozen adverbial units[1]. Adverbials have the advantage to be units for which linguists can easily wear a judgment of acceptability. Moreover, the proportion of specialized phraseologies is lower compared to noun or even verb phrases[2]. On the other hand, the limit between free and fixed sequence is hard to establish for adverbials, as fixed expressions are difficult to differentiate from collocations.

We decided to first concentrate on adverbials because these units seemed easier to tackle. Moreover, adverbials are central units in verbal polysemy processing and are acknowledged to be crucial for language description (Guimier, 1996; Molinier & Greyhound,

[1] This database was created by the FixISS team of the LDI Laboratory.

[2] Specialized vocabularies and terminologies are mostly described with verbs or nouns.

2000). They are predicate actualizers which are fundamental to describe predicates[3], enabling us to characterize and distinguish between predicates. All forms of disjoined adverbials are gathered under the name of *adverbial units* (Gross, G., 1988).

These units are extremely difficult to process as they are as heterogeneous as numerous. The number of compound adverbs is much greater than the number of simple adverbs. The most complete inventory for French has been established by M. Gross. In his book *La syntaxe de l'adverbe* (1986), the author draws parallels between the construction of fixed units and free syntax, and he assumes that the description of fixed units belongs to general syntax description. In that respect, batteries of tests specific to the described polylexical units should be developed. Syntactic regularities are often blurred by traditional grammars, which are based on heterogeneous criteria that do not distinguish between the different levels of analysis.

The syntax of adverbials would certainly deserve a more detailed analysis than can be given here. However, the classification we propose has the advantage to present a framework illustrating the diversity of adverbial units. We will resort to two major syntactico-semantic properties: (i) first, the units may be predicative or not predicative (Gross, G., 1988: 108). According to this criterion, we were able to distinguish between:

- binding adverbial units, which are beyond the sentence level. They are not part of the argumental scheme that characterizes them. These units carry various information, including:
  - field: *du point de vue philosophique* ('*from the philosophical point of view'*)
  - enunciation: *de toute évidence* ('*obviously'*)
  - reformulation: *autrement dit, en d'autres termes* ('*in other words'*)
  - discourse organizers: *en premier lieu, d'entrée de jeu* ('*first', 'from the very start'*)
  - connection: *en conséquence, au demeurant* ('*accordingly', 'as it happens'*)

- adverbial units that are second-order predicates, whose arguments are first other predicates. These adverbs are usually called "adverbs of manner" Considering them as predicates enables us to avoid choosing between "manner" or "means", which is often hard to determine:

  Luc marche (rapidement, de façon rapide)
  'Luke walks (quickly, in a quick way)'

  These phrases are not actualized themselves. The information is supported by the predicate of the main clause.

(ii) among adverbials, some units may have a wide scope, since they apply to a large number of predicates. Thus, the phrase *de manière correcte* ('correctly') may characterize all action predicates:

> *Luc a écrit ce mot de manière correcte* ('*Luke wrote the word correctly'*)
> *Luc s'est comporté de manière correcte* ('*Luke behaved correctly'*)

On the other hand, some phrases can only be used with a limited number of predicates (sometimes only one):

> *Luc crie/chante à tue-tête ('Luke shouts / sings very loud'*)
> *Léa est tombée sur son ennemi à bras raccourcis* ('*Lea set on his enemy'*)

The general category of "manner adverbs" has diverted researchers' attention away from the description of distributional restrictions characterizing the relation of certain adverbs to verbs. These adverbs can help differentiating action modes (*parler à voix haute/basse, à mi-voix… / 'speak out loud / low, softly...*') or states (eg. *se porter comme un charme / 'to feel as fit in a fiddle'*). These coocurrence restrictions have been described in detail by M. Gross (1986) who claimed that frozenness should be described by syntax and may take very different forms.

The current database has been compiled from books and dictionaries including:

- *La syntaxe de l'adverbe* (1986) from Maurice Gross, who has established the most comprehensive adverb typology for French. It includes 16 types of constructions (which are distributed into 6400 adverbial) and reflects the complexity of the phenomenon.

- The TLFi:
  The TLF (*Trésor de la Langue Française*) is a dictionary of the XIXe and XXe centuries in 16 volumes and one supplement. The TLFi is the computerized version of the TLF. It includes 100,000 words, 270,000 definitions and 430,000 examples. The TLFi differs from other existing electronic dictionaries by the precise way data are structured. A simple interface provides three levels of consultation: simple, assisted and complex searches. The user can enjoy all the resources of the TLF: searching for a word with approximate spelling, sorting of the different meanings, searching for a word within a given disciplinary field, searching for a sequence containing a particular part of speech, etc.

[3] Adverbials are indeed fundamental to describe verbal polysemy. Verb classes have to be described using appropriate adverbials (Grezka, 2009; Grezka & Martin-Berthet (eds), 2007). For instance, speech verbs are associated with various expressions describing noise level: *à voix haute/basse* ('out loud' / 'in a low voice', *à mi-voix* ('in an undertone'), etc. unless that characterization is already included in the verb (crier, gueuler, murmurer, etc. / 'to shout, scream, whisper, etc.'). Verbs of swallowing can also be modulated by amount and / or speed indications (*à petites gorgées, cul sec, etc.* / 'in small sips', 'bottoms up', etc.).

- *Le Petit Robert* CD Rom:
  The dictionary includes 60,000 words, 300,000 meanings, all verb tenses, etymology information, phonetic transcription, expressions and quotations; 1620 boxes on word etymological families; 16,000 words that are difficult to pronounce; 15,000 compounds with research index; and a total hypertext navigation.

2.2 Format and linguistic description

Users can currently search the database using two types of search: (i) headword search (1811 out of the 5000 entries, ie. units that are salient in the adverbial phrase, with a specific taxonomic role)[4] and (ii) pattern search.

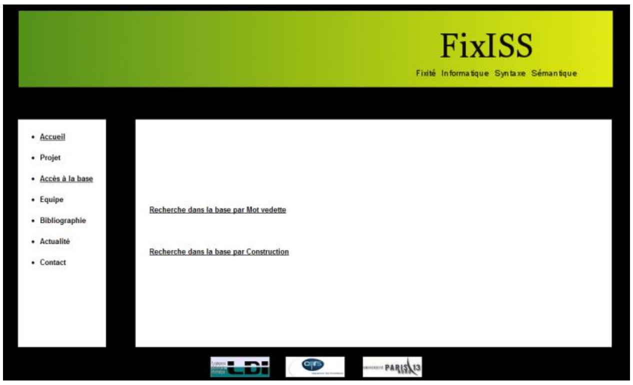
FixISS
Fixité Informatique Syntaxe Sémantique
Accueil
Projet
Accès à la base
Equipe
Bibliographie
Actualité
Contact
Recherche dans la base par Mot vedette
Recherche dans la base par Construction
PARIS 13

Figure 1: Database Home Page

By entering a word in the search engine, users can for instance consult:

- all the adverbial phrases that can be built using the entered word (for instance, *coup*):

  d'un coup d'aile
  d'un coup de baguette magique
  par un coup du sort
  à coups de pied au cul
  sans coup férir
  etc.

- all the possible lexical compositions of the word within the adverbial phrase, with possible expanded names (N of N, N adj, etc.):
  - *comme une fleur*: the entry is described by the keyword *fleur*;
  - *par un coup de poker*: the entry is first described by the keyword *coup* but another colum has been added containing the N of N *coup de poker*.

- all the associated constructions containing the word. These constructions are expressed using a systematic morphosyntactic notation.

A scientific description of frozenness requires the establishment of a typology of composition. Note that the elements of the description may vary, depending on the internal structure of adverbial units.

Entry	Lexical composition	Construction
à coups de pied au cul	coup de pied	Prép N Prép N Prép Dét N
à coups de pied au derrière	coup de pied	Prép N Prép N Prép Dét N
à coups de pied dans le cul	coup de pied	Prép N Prép N Prép Dét N
à coups de pied dans le derrière	coup de pied	Prép N Prép N Prép Dét N
au premier coup d'œil	coup d'œil	Prép Dét Ord N Prép N
d'un coup de baguette magique	coup de baguette magique	Prép Dét N Prép N Adj
par un coup de baguette magique	coup de baguette magique	Prép Dét N Prép N Adj
par un coup du sort	coup du sort	Prép Dét N Prép Dét N
pour le coup		Prép Dét N
sur le coup		Prép Dét N
sur un coup de dé	coup de dé	Prép Dét N Prép N
sur un coup de poker	coup de poker	Prép Dét N Prép N
sur un coup de tête	coup de tête	Prép Dét N Prép N

Table 1: Searching the word *coup ('hit')*

Most French grammars limit the description of adverbial units to a small number of morphological classes, such as: PREPOSITION + NAME (*par erreur, en catastrophe, de naissance*), PREPOSITION + DET + NAME (*contre toute attente, à la pièce, sous le manteau*), etc.
The necessity of a general typology has long been recognized by various authors. Precise analyses have enabled us to show the great diversity of adverbials in general language (Gross, 1996). Our research on adverbial units is an extension of this work. The systematic work we are conducting enables us to consider both the most original and complex phrases. The new typology that we propose will both be applied to general language and specialized languages. The patterns we found out are very diverse (more than 400 types of phrases already identified) and they will enable us to account for the morphological structure of the units.

- Finally, users can search for possible variations corresponding to series of tests measuring the frozenness degree of the units: deletion, move, change, etc. The tests we used are standard in the literature.

[4] It is besides interesting to underline that the most recurrent domains for these phrases are related to time and duration.

According to their needs, users can concentrate on one element or the other. The search can be oriented both on the internal structure of the phrase (and its lexical representation) and its external structure.

3. Frozeness criteria: linguistic tests

As with other grammatical categories, there is a continuum between free adverbial phrases and completely frozen ones. Frozenness is not an absolute phenomenon. The interest of its study is to trace the continuum, to describe frozenness degrees and to provide a description that will focus on the phenomena involved in the linguistic behavior of the frozen adverbial phrases. It is necessary to develop specific criteria to decide which of the possible units will be included into the database and to assess its frozenness degree. The establishment of these criteria has been extensively studied (Gougenheim, 1971; Gross, M., 1982; Mejri, 1997) and although the criteria do not have the same scope, they all state that structures are all the more frozen considering that they have fewer structuring and variation properties.

In order to assess the frozenness gradation of polylexical sequences, we need to perform general experiments shared by all sequences as well as experiments specific to the part of speech the sequence belongs to. The intersection of the two experiments will allow us to measure the frozenness degree of the unit. Many adverbial phrases accept variations on one of the elements and in the current state of the project, we can notably observe:

- Preposition variation :

  Nous avons rejoint Paris (de, en) une seule traite
  Tu y entres à coups de pied (dans, à) le cul

- Determination variation, with a rather open distribution (zero article, generic definite article, indefinite, partitive, possessive, numeral articles, etc.):

  Il a réparé l'appareil en (deux, trois, cinq) coups de cuiller à pot

  It should be noted that there are many constraints in determination. As for verbal phrases, adverbial units cannot be used with all persons. For example, phrases including a possessive adjective must co-refer to the subject (*de (ma, ta, sa) plus belle plume*):

  *Il avait écrit cette lettre de (sa/*ta) plus belle plume*
  *De (ma/*ta) plus belle plume je t'écris cette lettre d'amour*

- Determination also plays an important part in the semantics of the terms. The two following sentences do not have the same meaning at all in spite of their syntactic similarity:

  Tenir un cigare entre les dents
  Grommeler, murmurer, parler, répondre entre ses dents

- Lexical variants :

  Il devait nous rencontrer à onze heures (sonnantes, tapantes)
  Elle l'a sortie de la pièce à coups de pied dans le (cul, derrière)[5]

- Deletions :

  La cérémonie s'est déroulée dans l'intimité (la plus stricte)
  Elle s'est mise à pleurer d'un (seul) coup

- Position change :

  Luc parlait à voix haute et intelligible/à haute et intelligible voix

- Morphological variation :

  Luc travaille par intermittence(s)

- Graphical variation :

  Il y a une récompense à la clé/clef
  Luc connait sont sujet à fond/donf

- Hyphen variation :

  J'aime manger les légumes à la croque(-)au(-)sel

- Abbreviations :

  Luc fait son travail au max/maximum

We assume that the frozenness degree can be calculated from the results of the tests carried out above. The less the unit vary the less frozen it is. We consider a sequence is frozen when none of its components expresses a choice. For instance, no substitution is possible for a sequence such as *à fond la caisse* :

**À fond la (la voiture, la bagnole)*
** À fond (ta, notre) caisse*

We have presented the current database of frozen adverbial phrases we built. As said, most data have been extracted from the existing paper and electronic dictionaries. Nevertheless, the coverage of the database needs to be assessed using corpora. In this perspective, we will resort to various quantitative methods to further explore and refine the frozenness degree of the units.

4. Using corpora and text statistics methods to assess frozenness and extend the database

The database we have presented supra aims to be exhaustive for the French language and will ultimately associate an index expressing a degree of frozenness to each unit, calculated from a battery of language tests relevant in the literature.

Although it is in its current state a useful object of consultation for linguists, the database is still free of

[5] This variation is due to register change.

indications concerning phrase *usage* in texts and discourses and this is a problem since we know that certain phrases are specific to certain genres and domains: this significantly impact their interpretation as well as their frozenness degree. A fortiori, the database is little operational with this limitation, as NLP and text mining applications run in targeted and restricted discursive usages.

The aim of the present section is to explore and assess the usefulness of different corpus processing methods to describe the database using usage indications. This research will notably enable us to uncover different possibilities of automatic extension of the database using corpora representative of various discursive usages.

We first took advantage of the most frequent morphosyntactic patterns encoded in the database. As shown in Figure 1, more than 50% of the frozen adverbials do not project from adverb heads but occur as prepositional phrases.

Morphosyntactic pattern	**Occ.**	**%**
Prep Det N	949	18,66%
Prep N	818	16,08%
Prep Adj N	280	5,51%
Prep Det Adj N	225	4,42%
Prep N Adj	204	4,01%
Prep Det N Adj	174	3,42%
Prep Det N Prep Det N	143	2,81%
Prep N Prep N	139	2,73%
		57,64%

Table 2: Most frequent patterns in the database

The patterns are highly ambiguous, since a prepositional phrase is of course far from being systematically an adverb. However, we concentrated on these structures in the following analyses to extract adverbial phrase candidates that might help determining a relevant frozenness degree, as well as to increase the base.

We resorted to three distinct corpora in this study, representing two types of discourse: political and scientific discourses (4.1.). We first used the morphosyntactic patterns to extract adverbial phrase candidates from the political and scientific corpora. The extractions were performed using concordances and indexes expressed in CQP with or without gap using the TXM opensource software (4.2.).

We will try to answer the following questions:

- Are the frozenness patterns we observed in the database more productive from one discourse to another? What are the privileged frozen constructions from one discourse to another? To what extent are they interesting for corpus exploration and description?
- To what extent can relative frozenness and defrozenness be considered using concordance and index functionalities?

4.1. Corpora

In that respect, we will resort to two different French corpora representing two types of discourse:

- **political discourse** (2 millions occ.) using a corpus developed by D. Mayaffre which collects the discourses of Vth Republic French Presidents, from de Gaulle to Sarkozy (1958-2010). It gathers 800 general public speeches, including about a hundred mandate speeches;

- **scientific discourse in linguistics** (2,4 millions occ., described in Poudat, 2006), which contains 224 articles published in 32 journals belonging to the field of linguistics.

The choice of these two corpora was motivated by the fact that political and scientific discourses are little used to build such databases; the FixISS database was notably built from a set of dictionaries which were themselves based on Frantext, that is, literary texts.

4.2. Multi-level concordance

Corpus linguistics has significantly influenced linguistic descriptions and practices. Corpus linguistics concentrates on naturally occurring data and designs corpora to investigate language use and variation, enabling researchers to assess distances between intuition and data as well as to explore large corpora. Numerous corpus tools have been developed to carry out analyses on the patterns found in natural texts. In this process, concordance programs are crucial, and in fact they are essentially the main tools used by corpus linguists. For example, when Baker (2010) lists the most popular corpus tools, he actually lists the major concordance programs (Wordsmith Tools[6], AntConc[7], MonoConc Pro[8], Xaira[9], Sketch Engine[10], COBUILD Concordance Sampler[11]). The search for patterns is indeed at the heart of corpus linguistics, as collocations are considered to be one of the two main organizing features of texts (see Sinclair's *idiom* and *open choice* principles, Sinclair 1991).

The tool we will resort to has been developed within the framework of text statistics, which is a particularly well developed field in France. This movement originated in Saint Cloud, France in the 1980s and from the beginning, it was a corpus-based approach, which built and contrasted corpora using statistical measures such as relative frequencies. In this framework, the norm is endogenous to the corpus, as linguistic units

[6] http://www.lexically.net/wordsmith/, accessed 12/03/12.

[7] http://www.antlab.sci.waseda.ac.jp/software.html, accessed 12/03/12.

[8] http://www.athel.com/mp.html, accessed 12/03/12.

[9] http://www.oucs.ox.ac.uk/rts/xaira/, accessed 12/03/12.

[10] http://www.sketchengine.co.uk/, accessed 12/03/12.

[11] http://www.harpercollins.co.uk/about-harpercollins/Imprints/collins/Pages/Collins.aspx, accessed 12/03/12.

do not have frequencies in Language (Lafon 1980). The methods developed were quite innovative for their time, and they are still frequently used for exploring and contrasting large corpora. Although they are not very well known outside of the French-speaking scientific community, the methods and software developed would be very useful for corpus linguists, who mainly work with concordance programs that include comparatively basic statistical measures.

A recent project[12] involving the major software designers (Hyperbase, Lexico3, Weblex[13]) has led to the development of a standalone open-source platform, TXM, which should be powerful enough to challenge the existing corpus tools. The TXM workbench notably includes an impressive concordance program that enables researchers to perform multi-level searches on morphosyntactically labeled corpora (i.e. words, lemmas and morphosyntactic categories, thanks to a Corpus Query Processor[14] that also allows users to submit regular expressions). Queries were performed using TXM on the political and scientific corpora, on the basis of the eight morphosyntactic patterns found in the database (as seen in Figure 2).

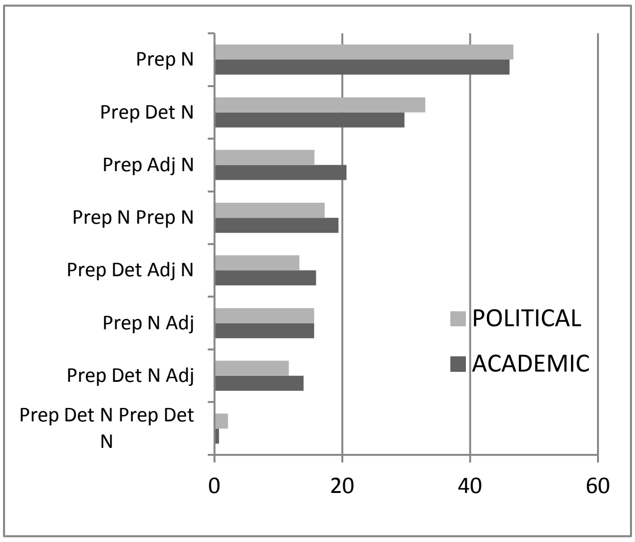

Figure 2. Database pattern assessment

Before presenting the results provided by the concordance, let us assess the database by observing the different percentages of the phrases identified in the two corpora. As we can see in Figure 2, the results are comparable from a corpus to another: we identified over 45% of the phrases following the pattern *Prep N* while the pattern *Prep Det N Prep Det* N proved to be the least operational for extraction.

If we now extract the morphosyntactic patterns from the two corpora to assess the database coverage, the results are naturally quite different: the adverbial phrases included in the database represent less than 2% of all sequences extracted (resp. 1,28% political corpus / 1,36% academic corpus). This low coverage is naturally correlated to the pattern ambiguity since the adverbial we are interested in do not project from adverb heads but occur as prepositional phrases, in all cases, filters are required to extract relevant phrases.

	POL - distinct patterns	AC - distinct patterns	POL - occ.	AC - occ.
Prep Det N	26 298	25 867	96 693	77 975
Prep N	9724	13 344	60 711	61 421
Prep Adj N	1737	1335	4220	2601
Prep Det Adj N	3770	3196	5732	4375
Prep N Adj	4581	7062	6598	9357
Prep Det N Adj	11 496	14 243	17 242	17 323
Prep Det N Prep Det N	7468	6345	8623	6804
Prep N Prep N	2809	3390	3552	4167
TOTAL	67 883	74 782	203 371	184 023

Table 3: Distribution of the eight patterns in the two corpora (without gap)

As the two corpora we resorted to are of comparable size (about 2 million words), raw figures are presented in Table 3; we distinguished between forms and total occurrences.

The nature of the patterns first vary from a corpus to another, with respect to the characteristics of the two discourses: academic texts contain more concepts and notions and this significantly increases the number of prepositions immediately followed by a noun (NP *de langue, de sens, de parole, d'interprétation, etc.*). The *Prep N Prep N* sequences (e.g. *d'acte de langage, d'univers de discours*) and *Prep N Adj* (e.g. *de linguistique générale, en mémoire discursive*) are also highly represented. This prevalence of specialized concepts and phraseologies also impacts the *Prep Det Adj N* sequence, with N Adj noun phrases (e.g. *langue parlée, langue française, forme schématique*).

Regarding unit extraction and filtering, and the determination of a frozenness degree, several tracks will be followed to isolate adverbial phrase candidates; criteria will be weighted and combined for greater efficiency:

- First of all, the most frequent nouns found out in the database should first be associated with a specific weight, ie *fois* (93 occ.), *heure* (88), *jour* (70), *moment* (69), *coup* (57), *main* (46), *œil* (41), *titre* (40), *manière* (40) and *pas* (37). Note that these terms are differently used in a discourse and another. For instance, political discourse rather uses

[12] See the Project website at http://textometrie.ens-lyon.fr/, accessed 12/03/12.

[13] Developed by Heiden: http://weblex.ens-lsh.fr/doc/weblex/, accessed 12/03/12.

[14] IMS Corpus Workbench (Christ et al. 1999): http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/, accessed 12/03/12.

the adverbial *à l’heure* (1279 vs 21 occ. in AC) if we consider the *Prep Det N* pattern while scientific discourse largely prefers *à la fois*, or *dans la mesure (où)* ;

- A close examination of the candidates highlights the interest of considering the relation between the frequency of the adverbial phrase and the frequency of the noun it contains in the corpus. Besides, this echoes the previous work we conducted on cooccurrences asymmetry (Luong et al., 2010): the attraction of a given noun for the phrase in which it occurs seems to be an interesting frozenness criteria. Nouns that give over 35% of their occurrences to a given pattern seem to point to an adverb, as shown in Table 4;

ACADEMIC DISCOURSE			
Adverbial frozen unit candidate	**Occ.**	**Occ. noun corpus**	**%**
de la **langue**	545	2075	26,27
*à la **fois**	375	839	44,70
de l'**énoncé**	336	1210	27,77
dans la **mesure**	306	605	50,58
*d'une **part**	300	1149	26,11
dans le **cadre**	282	637	44,27
de la **phrase**	273	1211	22,54
dans le **cas**	245	2578	9,50
à l'**intérieur**	215	525	40,95
de la **relation**	193	1680	11,49
de la **forme**	189	1907	9,91
à l'**aide**	129	176	73,30
en l'**occurrence**	107	283	37,81
POLITICAL DISCOURSE			
Adverbial frozen unit candidate	**Occ.**	**Occ. noun corpus**	**%**
de la **France**	1765	7753	22,77
*à l'**heure**	1279	1510	84,70
de la **République**	1130	1676	67,42
sur le **plan**	541	1064	50,85
*à la **fois**	536	2578	20,79
pour la **France**	347	7753	4,48
dans le **cadre**	274	443	61,85
à l'**intérieur**	271	403	67,25
à l'**égard**	266	556	47,84

Table 4: Ratio noun / adverbial phrases – pattern Prep Det N (* already included in the database)

- This observation must obviously be completed by considering the specificity scores of the nouns from a corpus and a discourse to another. Thus, *cas* (case) is an object handled by linguistics and this hampers the identification of *dans le cas* (in the case). TXM notably computes ‘specificities’, which are calculated according to a probabilistic model (Lafon 1980) based on hypergeometric distributions – they share the same goals as the keywords facility of Wordsmith for instance[15];

- The presence of gaps between the pattern units is also a relevant indicator to assess the frozenness of an adverbial. CQL allows the presence of gaps, whose number is defined by the user. We have launched a second series of queries with gaps of 1 to 3 elements, which allowed us to extract adverbial phrases that were not referenced in the database. Here are some examples of the motifs we extracted:

    - **Prep Det N** : *avec relation du tout à la partie* (8 occ. AC), *dans toute la mesure du possible* (7 occ. POL), *à partir d’un certain moment* (6 occ. POL), *d’un bout à l’autre du monde* (5 occ. POL) / *de l’Europe* (4 occ. POL) ;

    - **Prep Adj N** : *dans l’état actuel des choses* (12 occ. POL), *à la troisième personne du singulier* (8 occ. AC), *dans un certain nombre de domaines* (7 occ. POL), *en désaccord total avec les prédictions* (7 occ. AC), *pour un certain nombre de raisons* (5 occ. POL), *dans un certain nombre de secteurs* (5 occ.) ;

    - **Prep N Adj** : *d’un point de vue sémantique* (11 AC), *diachronique, cognitif, pragmatique, syntaxique* (4 AC), *de la façon la plus claire* (7 POL) / *nette* (6 POL) ;

    - **Prep N Prep N** : *dans le cadre de la théorie* (12 AC), *dans l’histoire de la République* (11 POL), *dans le cadre d’une théorie* (6 AC), *dans la perspective d’une qualification* (5 AC).

Note that these elements are also informative to describe the rhetoric and the fixity of texts and discourses.

5. Conclusion and perspectives

We have presented the FixIss database and exposed a set of possible methods that would enable us to extend and assess the database and its coverage. Among the methods presented, queries with gaps were particularly interesting to determine a frozenness degree.

We are currently testing a new approach which allows us to detect multilevel motifs in an entirely inductive way. This inductive experimental methodology has been developed in Caen, France, in the framework of a collaboration between the GREYC and the CRISCO labs. The approach is described in (Quiniou et al. 2012a, 2012b) and goes beyond measures such as Salem’s *repeated segments* (which are implemented in the Lexico3 software).

We applied the algorithm on the same data sets and the motifs we found out were significantly different. For instance, motifs following the morphosyntactic pattern *PREP N PREP N* differ from one corpus to another:

[15] http://www.lexically.net/wordsmith/, accessed 12/03/12.

- POL: *pour* N *pour* N : *pour la France et pour les Français* (6 occ.), *pour l' Europe et pour la France* (4 occ.), *pour le progrès et pour la paix* (4 occ.), *pour l' Europe , pour le monde*, etc.

- AC : *de* N *de* N : *de la métaphore et de la métonymie* (8 occ.), *de la syntaxe et de la sémantique* (4 occ.), *de la métonymie et de la métaphore* (4 occ.), *de son sujet ou de son objet* (3 occ.), etc.

This provides relevant research directions to enrich the frozenness database with usage information (following the example of the Longman grammar designed by Biber et al., 1999) and will help us determining in the long term a frozenness degree combining linguistic tests and statistical measures.

6. References

Baker, P. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh University Press, Edinburgh.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (Eds.), 1999, *Longman Grammar of Spoken and Written English*. London: Longman.

Brunet, E., 2011, *Ce qui compte. Écrits choisis, tome II – Méthodes statistiques*. Textes édités par Céline Poudat. Préface de Ludovic Lebart. Collection *Lettres numériques*, vol. 11, Honoré Champion, Paris.

François, J. et S. Mejri (eds.), 2006, *Composition syntaxique et figement lexical*, Coll. Bibliothèque de Syntaxe & Sémantique. Presses Universitaires de Caen, 280p.

Gougenheim, G., 1971, *Étude sur les périphrases verbales de la langue française*. Nizet, Paris, 383p.

Grezka, A., 2009, *La polysémie des verbes de perception visuelle*. Collection *Sémantiques*, L'Harmattan, Paris, 292p.

Grezka, A. et F. Martin-Berthet (eds), 2007, "Verbes et classes sémantiques", *Verbum*, 2007:1, Presses Universitaires de Nancy, 172p.

Gross, G., 1996, *Les expressions figées en français. Noms composés et autres locutions*. L'essentiel français. Paris, Ophrys, 161p.

Gross, G., 1988, « Degré de figement des noms composés », *Langages* N°90, pp. 57-70.

Gross, M., 1986, *Grammaire transformationnelle du français. 3. Syntaxe de l'adverbe*, ASSTRIL, Paris.

Gross, M., 1982, « Une classification des phrases figées du français », *Revue Québécoise de Linguistique*, 11(2), pp. 151-185.

Grossmann, F. et A. Tutin, 2003, *Les collocations : analyse et traitement*, Amsterdam, De Werelt, 144p.

Guimier, C., 1996, *Les adverbes du français. Le cas des adverbes en –ment*, L'essentiel français. Paris, Ophrys, 170p.

Lafon P., 1980, « Sur la variabilité de la fréquence des formes dans un corpus », *Mots*, p 127-165.

Luong, X., Brunet, E., Longrée, D., Mayaffre, D., Mellet, S., & Poudat, C. (2010). La cooccurrence, une relation asymétrique? *in* Bolasco, Chiari, Giuliano (éds.), Actes des *JADT 2010* - In S., Bolasco, I., Chiari, & L., Giuliano (Eds.), *Statistical Analysis of Textual Data: Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles 9-11 June 2010 - Sapienza University of Rome* (pp. 321-331). Milan, Italie: Edizioni Universitarie di Lettere Economia Diritto.

Mejri, S., 2009, « Le mot, problématique théorique ». *Le Français Moderne* 77 (1), pp. 68-82.

Mejri, S., 2005, « Figement absolu ou relatif : la notion de degré de figement ». *Linx* N°53, Université Paris X Nanterre, pp.183-196.

Mejri, S., 2003, « Le figement lexical ». *Cahiers de Lexicologie* 82, pp. 23-39.

Mejri, S., 1997, *Le figement lexical : descriptions linguistiques et structuration sémantique*, Tunisie, Publications de la Faculté de lettres de la Manouba, 632p.

Longrée D., Mellet S. 2009. Syntactical Motifs and Textual Structures, *Belgian Journal of Linguistics*, 23 ("*New Approaches in Textual Linguistics*"), pp. 161-173.

Molinier, C. et F. Levrier, 2000, *Grammaire des adverbes. Description des formes en –ment*, Droz, Genève-Paris.

Poudat, C., 2006. « Étude contrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres » *in Texto!* [en ligne], septembre-décembre 2006, vol. XI, n°3-4. Disponible sur http://www.revue-texto.net/1996-2007/Corpus/Corpus.html

Quiniou, S., Cellier, P., Charnois, T et Legallois, D. (2012a) « Fouille de données pour la stylistique : cas des motifs séquentiels émergents ». JADT 2012, Liège, Belgium (forthcoming).

Quiniou, S., Cellier, P., Charnois, T et Legallois, D. (2012b) "What About Sequential Data Mining Techniques to Identify Linguistic Patterns for Stylistics ?". CICLING 2012, Lectures Notes in Computer Sciences, Springer Verlag.

Sinclair 1991. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Svensson, M.-H., 2004, *Critères de figement. L'identification des expressions figées en français*, Umeå, Umeå universitet.