

An Electronic Dictionary of French Multiword Adverbs

Éric Laporte, Stavroula Voyatzi

Université Paris -Est

IGM-Labinfo

5, Boulevard Descartes, Champs-sur-Marne

77454 Marne-la-Vallée Cedex 2 (France)

E-mail: eric.laporte@univ-paris-est.fr, voyatzi@univ-mlv.fr

Abstract

We present an electronic dictionary of French multiword adverbs. This dictionary is designed for investigation on information retrieval and extraction, automatic lexical acquisition, as well as on deep and shallow syntactic parsing. We delimit the scope of the dictionary in terms of lexical coverage and of grammatical coverage, we outline the formal description of entries, and we give an overview of the syntactic and semantic features which are associated to the 6,800 adverbial entries of the lexicon. This electronic dictionary is freely available on the web.

1. Introduction

Recognising multiword adverbs such as *à long terme* 'in the long run' in texts is likely to be useful for information retrieval and extraction because of the information that some of these adverbials convey. In addition, it is likely to help resolving prepositional attachment during shallow or deep parsing: most multiword adverbs have the superficial syntax of prepositional phrases; in many cases, recognising them rules out attachments where they are analysed as arguments or noun modifiers.

In the current practices of natural language processing, the handling of multiword expressions (MWEs) in general is in its infancy. Much research effort towards MWE recognition is devoted to algorithms, but results depend also on resources. We describe an electronic dictionary of multiword adverbs of French with syntactic-semantic information. This dictionary is freely available on the web under LGPL license. In this article, we survey related work, we define the scope of the dictionary, we present the syntactic and semantic features assigned to entries and we describe their representation.

2. Related research

A considerable amount of research has been conducted in the area of MWEs, e.g. general studies (Sag *et al.*, 2002) and efforts towards standardization (Calzolari *et al.*, 2002), but they seldom rely on large-coverage lexical resources¹. Michiels and Dufour, (1998) exploit conventional dictionaries (i.e. written for human readers), but such resources have well-known inherent limitations. However, there do exist NLP-oriented lexicons with a large coverage in MWEs, including multiword adverbs, e.g. WordNet

(Miller, 1995). Lexicological research focusing on multi-word adverbs has been devoted to French (Gross, 1990), German (Seelbach, 1990), Spanish (Blanco & Català, 1998/1999), Italian (De Gioia, 2001), Portuguese (Baptista, 2003), Korean (Jung, 2005) and Modern Greek (Voyatzi, 2006) with the Lexicon-Grammar methods of NLP-oriented lexicon design (Gross, 1986; 1994), on the basis of conventional dictionaries, grammars, corpora and introspection². Català and Baptista (2007) show that multiword adverbs are recognized in Spanish text with 77% precision through the use of a Lexicon-Grammar.

In parallel, research on automatic lexical acquisition was targeted both at terminology (Daille, 2000) and general-language MWEs. Such techniques use both statistical approaches and linguistic information, such as parts of speech and inflectional categories, and require large corpora that contain significant numbers of occurrences of MWEs. However, even with corpora of millions of words, frequencies of MWEs are usually too low for statistical extraction (Mota *et al.*, 2004). Gross (1986) reports that the number of MWEs in the lexicon of a language is larger than the number of single words (cf. also Jackendoff, 1997), therefore any extraction method must be able to handle extremely sparse data. In addition, adverbs or more generally non-object complements have not been the focus of attention, and their relations to simple sentences are far from being understood³.

² The resulting resources on French, enclosed in the Intex system (Silberztein, 1994), have helped to annotate the French Treebank (Abeillé *et al.*, 2003), in which prepositional phrases and adverbs are annotated with a binary feature ('compound') which indicates whether they are multiword units; the distinction between whether prepositional phrases are verb modifiers, noun modifiers or objects appears only in the function-annotated part of the Treebank (350,000 words).

³ Several reasons explain this lack of interest. Firstly, adverbials are usually felt as less useful than nouns for information retrieval and extraction. Secondly, many multiword adverbs are difficult to distinguish from prepositional phrases assuming other syntactic functions, such as arguments or noun modifiers: the distinction is hardly correlated to any material markers in texts

¹ In practice, paradoxically, even investigation in semi-automatic extension of MWE lexicons pays little attention to the structure and contents of existing large-coverage lexicons (e.g. Navigli, 2005). Copestake *et al.* (2002) contains interesting thoughts, but they are not validated against an available large-coverage lexicon and it does not deal with adverbials.

The availability of large-coverage lexicons of multiword adverbs is essential to gaining insight on their recognition, including the dual problems of variability and ambiguity. The resource described in this paper is the Lexicon-Grammar of French multiword adverbs (Gross, 1990), in which previously implicit features have been made explicit for more convenient use in NLP.

3. Scope of lexicon

The scope of the lexicon is delimited by the intersection of two criteria: (i) multiword expressions and (ii) adverbial function. In this section, we define both criteria in more detail and we present the features provided in the lexicon.

3.1 The multiword unit criterion

For this work, a phrase composed of several words is considered to be a multiword expression if some or all of its elements are frozen together, that is, if their combination does not obey productive rules of syntactic and semantic compositionality. In the following example, *de nos jours* ('nowadays', lit. 'of our days') is a multiword unit assuming an adverbial function:

- (1) *Il est facile de nos jours de s'informer*
'It is easy to get informed **nowadays**'

This criterion ensures a complementarity between lexicon and grammar. In other words, it tends to ensure that any combination of linguistic elements which is licit in the language, but is not represented in syntactic-semantic grammars, will be stored in lexicons.

Syntactic-semantic compositionality is usually defined as follows: a combination of linguistic elements is compositional if and only if its meaning can be computed from its elements. This is also our conception. However, in this definition, we consider that the possibility of computing the meaning of phrases from their elements is of any interest only if it is a better solution than storing the same phrases in lexicons, i.e. if they rely on grammatical rules with sufficient generality. In other words, we consider a combination of linguistic elements to be compositional if and only if its meaning can be computed from its elements **by a grammar**. In example (1) above, the lack of compositionality is apparent from distributional restrictions⁴ such as:

- * *Il est facile de nos semaines de s'informer*
* 'It is easy to get informed nowa**weeks**'

and by the impossibility of inserting modifiers that are a priori plausible, syntactically and semantically:

- * *de nos jours (de repos + de fête)*
literally 'of our days (of rest + of feast)'

and lies in complex linguistic notions (Villavicencio, 2002; Merlo, 2003).

⁴ The point is that this blocking of distributional variation (as well as other syntactic constraints) cannot be predicted on the basis of general grammar rules and independently needed lexical entries. Therefore, the acceptable combinations are meaning units and have to be included in lexicons as multiword lexical items.

- pendant nos jours (de repos + de fête)*
literally 'during our days (of rest + of feast)'

MWEs include many different subtypes, varying from entirely fixed expressions to syntactically more flexible expressions (Sag *et al.*, 2002). In (2), the possessive adjective agrees obligatorily in person and number with the subject of the sentence:

- (2) *De (ses + *mes) propres mains, il a construit une maison en torchis*
'**With (his + *my) own hands**, he built a house in cob'

The lexicon also takes into account expressions which comprise a frozen part and a free part, e.g. *au moyen de ce bouton* 'with the aid of this switch'. The frozen part *au moyen de* 'with the aid of' is encoded in the lexicon, and the syntactic category of the free part, here *NP*, is encoded as a feature⁵. Open classes of multiword adverbs such as named entities (NEs) of date or duration are not included in the dictionary, since they follow quite specific syntactical rules and use a closed lexicon. They can be identified with FST methods (Martineau *et al.*, 2007).

3.2 The adverbial function criterion

The dictionary deals only with MWEs which can assume an adverbial role, i.e. circumstantial complements, or complements which are not objects of the predicate of the clause in which they appear. They are identified through criteria (Gross, 1986; 1990) involving the fact that they are optional, they combine freely with a wide variety of predicates and some of them pronominalize with specific forms. Phrases with adverbial function are often called 'circumstantial complements', 'adverbials', 'adjuncts', or 'generalised adverbs'. They assume several morphosyntactic forms: underived (*demain* 'tomorrow') or derived adverbs (*prochainement* 'soon'), prepositional phrases (*à la dernière minute* 'at the last minute') or circumstantial clauses (*jusqu'à ce que mort s'ensuive* 'until death comes'), and special structures in the case of NEs of time (*lundi 20* 'on Monday 20') (cf. section 3.1).

3.3 The features

French multiword adverbs have been assigned a feature describing their internal morphosyntactic structure. The definition of the morphosyntactic structures is based on the number, category and position of the frozen and free components of the adverbial. They are described as a sequence of parts of speech and syntactic categories. For example, *à la nuit tombante* 'at nightfall' is assigned a structure identified by the mnemonic acronym *PCA*, and defined as *Prép Dét C (MPA) Adj*, where *C* stands for a noun frozen with the rest of the adverbial, *Adj* for a post-posed noun modifier (e.g. an adjectival phrase or a relative clause), and *MPA* for a pre-adjectival modifier, empty in this lexical item. The 15 structures, together with an illustrative example and the corresponding number of entries are listed in Table 1.

⁵ In case of a limited set of possibilities, all of them are listed in independent entries, as in *au sens propre (du mot + du terme + de l'expression)* 'in the proper sense of the (word + term + phrase)'.

Struct.	Example	English equivalent	Size
PC	<i>par exemple</i>	for example	664
PDETC	<i>de nos jours</i>	nowadays	848
PAC	<i>à la dernière minute</i>	at the last minute	776
PCA	<i>à la nuit tombante</i>	at nightfall	840
PCDC	<i>dans la limite du possible</i>	as far as possible	750
PCPC	<i>à cent pour cent</i>	one hundred percent	287
PCONJ	<i>tôt ou tard</i>	sooner or later	333
PCDN	<i>à l'insu de NP</i>	unbeknowst to NP	555
PCPN	<i>en comparaison avec NP</i>	in comparison with NP	151
PV	<i>à dire vrai</i>	to tell the truth	285
PF	<i>jusqu'à ce que mort s'ensuive</i>	until death comes	396
PECO	<i><fidèle> comme un chien</i>	as <faithful> as a dog	305
PVCO	<i><travailler> comme un chien</i>	<work> as much as a dog	338
PPCO	<i><disparaître> comme par enchantement</i>	<vanish> as by enchantement	50
PJC	<i>mais aussi et surtout</i>	but also and foremost	185
Total			6,763

Table 1: Morphosyntactic structures of multiword adverbs

Examples of other syntactic-semantic features provided in the lexicon are (i) the conjunctive function of the adverbial in discourse, (ii) the omission of the pre-adjectival modifier *MPA* without loss of information, or (iii) the constraint that the adverbial obligatorily occurs in a negative clause (cf. section 4 and table 2).

4. The Electronic Dictionary

The electronic dictionary of French multiword adverbs has 6,800 entries. It is freely available⁶ for research and business under the LGPL license. It takes the form of a set of Lexicon-Grammar tables (or binary matrices) such as that of Table 2, which displays a sample of the lexical items with the *PCA* morphosyntactic structure.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	NO = Nhum	NO = N-hum	Nég obl											
				Prép	Dét	C		Modif pré-adj		Adj			Prép Dét C	Prép Dét MPA Adj C
													Prép MPA Adj C	Conjunction
170	+	-	-	agr	dans	les	délais	les plus	brefs					
171	+	-	-	agr	dans	les	délais	les plus	courts					
172	+	-	-	agr	dans	les	délais	les meilleurs						
173	+	-	-	rtr	<E>	toutes	dents	<E>	dehors					
174	-	+	-	se produire	à	cette	époque-	<E>	ci					
175	-	+	-	se produire	à	cette	époque-	<E>	là					

Table 2: Sample of the table of entries with the *PCA* morphosyntactic structure

In this table, each row describes a lexical item, and each column corresponds:

- either to one of the elements in the morphosyntactic structure of the items (columns with identifiers 'Prép', 'Dét', 'C', 'Modif pré-adj' and 'Adj');

⁶ <http://infolingu.univ-mlv.fr/english/DonneesLinguistiques/Lexiques-Grammaires/View.html>.

- or to a syntactic-semantic feature (cf. 3.3); these columns hold binary values: 'Conjunction', 'Prép Dét C', 'Nég obl';

- or to illustrative information provided as an aid for the human reader to find examples of sentences containing the adverbial (e.g. columns D and E giving an example of a verb compatible with the adverb).

There are 15 such tables, one for each of the morphosyntactic structures.

4.1 The General Table

A lexicon is not a static resource: it has to be updated with the evolution of language. In order to facilitate the manual maintenance of the lexicon by linguists, the following organization has been adopted. When the values of a syntactic or semantic feature are the same over all entries in a class, it is not displayed in the corresponding class table. We stored it in a General Table (Figure 3).

	A	J	K	L	M	N	D	P	Q	R	E	T	U	V	X	Y	Z	AA	AB	AC			
1	Table	Prép2	Dét2	C2	Conj	Ppv	V	ConjS	Dét0	C0	comme	de Dind façon C-a	C-lement	C-a	Conjunction	Prép Dét C	Prép Dét MPA Adj C	Prép MPA Adj C	Prép1 Dét1 C1	N2=Nhum	N2=N-hum	C1 de N2=Poss2 C1	
2	PC																						
3	PDETC																						
4	PAC																						
5	PCA																						
6	PCDC	de	o																				
7	PCPC	o	o	o	o																		
8	PCONJ	o	o	o	o																		
9	PCDN	de																					
10	PCPN																						
11	PV																						
12	PF																						
13	PECO										comme												
14	PVCO										comme												
15	PPCO										comme												
16	PJC																						

Figure 3: Sample of General Table of multiword adverbs

The rows correspond to the morphosyntactic structures of multiword adverbs. All the 29 features described in any of the 15 tables are represented in the columns of the general table. Moreover, it also takes into account 12 features that had not been encoded in any of the 15 tables (for example, features connected with the morphosyntactic structures), totaling 41 features, all described in our documentation available with the lexicon. Values used at the intersection of rows and columns indicate that the feature in the column:

- is encoded in the table associated to the row; its value is variable (noted 'o');
- is encoded in the table associated to the row; its value is constant (noted <value>, e.g. Prép2='de');
- is not encoded in the table associated to the row; if it were encoded, its value would be constant (noted <value>, e.g. '+' or '-');
- is not encoded in the table associated to the row; if it were encoded, its value would be variable (noted 'O').

5. Conclusion

This paper described the design of an electronic lexicon of

French multiword adverbs which comprise 6,800 fixed, semi-flexible and flexible combinations, all of them associated with appropriate morphosyntactic and semantic features. This electronic dictionary is freely available on the web for research on information retrieval and extraction, automatic lexical acquisition, as well as on deep and shallow syntactic parsing.

6. Acknowledgements

This task has been partially financed by CNRS and by regional business cluster Cap Digital.

7. References

- Abeillé, A., Clément, L., and Toussnel, F. (2003). Building a Treebank for French. In A. Abeillé (Ed.), *Building and Using Parsed Corpora, Text, Speech and Language Technology*, 20, Kluwer, pp. 165–187.
- Baptista, J. (2003). Some Families of Compound Temporal Adverbs in Portuguese. In *Proceedings of the Workshop on Finite-State Methods for Natural Language Processing, EACL*, Budapest, pp. 97–104.
- Blanco, X., Català, D. (1998/1999). Quelques remarques sur un dictionnaire électronique d'adverbes composés en espagnol. *Lingvisticae Investigationes* 22, pp. 213–232.
- Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., MacLeod, C. and Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC 2002*, Las Palmas, pp. 1934–1940.
- Català, D., Baptista, J. (2007). Spanish Adverbial Frozen Expressions. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions, ACL 2007*, Prague, Czech Republic, pp. 33–40.
- Copestake, A., Lambeau, F., Villavicencio, A., Bond, F., Baldwin, T., Sag, I., Flickinger, D. (2002). Multiword Expressions: Linguistic Precision and Reusability, In *Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC 2002*, Las Palmas, pp. 1941–1947.
- Daille, B. (2000). Morphological rule induction for terminology acquisition. In *Proceedings of the 18th International Conference on Computational Linguistics, COLING'00*, Saarbrücken, Germany, pp. 215–221.
- De Gioia, M. (2001). *Avverbi idiomatici dell'italiano. Analisi lessico-grammaticale, prefazione di Maurice Gross*, Torino, L'Harmattan.
- Gross, M. (1986). Lexicon-Grammar. The representation of compound words. In *Proceedings of the 11th International Conference on Computational Linguistics, COLING'86*, Bonn, West Germany, pp. 1–6.
- Gross, M. (1990). *Grammaire transformationnelle du français: 3. Syntaxe de l'adverbe*. Paris, ASSTRIL.
- Gross, M. (1994). Constructing Lexicon-Grammars. In Atkins & Zampolli (Eds.), *Computational Approaches to the Lexicon*, Oxford University Press, pp. 213–263.
- Jackendoff, R. (1997). *The architecture of the Language Faculty*. Cambridge, MA, MIT Press.
- Jung, E. J. (2005). *Grammaire des adverbes de durée et de date en coréen*. Thèse de doctorat en Informatique Linguistique. Université Paris -Est Marne-la-Vallée.
- Martineau, C., Tolone, E., Voyatzi, S. (2007). Les Entités Nommées : usage et degrés de précision et de désambiguïsation. In *Proceedings of the 26th International Conference on Lexis and Grammar*, Bonifacio, pp. 105–112.
- Merlo, P. (2003). Generalised PP-attachment Disambiguation using Corpus-based Linguistic Diagnostics. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, EACL*, Budapest, pp. 251–258.
- Michiels, A. and Dufour, N. (1998). DEFI, a tool for automatic multi-word unit recognition, meaning assignment and translation selection. In *Proceedings of the 1st International Conference on Language Resources and Evaluation, LREC*, Granada, pp. 1179–1186.
- Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38:11, pp. 39–41.
- Mota, C. Carvalho, P. Ranchhod, E. (2004). Multiword Lexical Acquisition and Dictionary Formalization. In Michael Zock (ed.), *In Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries, COLING 04*, Geneva, pp. 73–76.
- Navigli, R. (2005). Semi-Automatic Extension of Large-Scale Linguistic Knowledge Bases, In *Proceedings of the Florida AI Research Society Conference*, pp.548–553.
- Sag, I. A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In A. Gelbuk (Ed.), *Computational Linguistics and Intelligent Text Processing: Proceedings of the Third International Conference, CICLing 2002*, Springer-Verlag, Heidelberg/Berlin, pp. 1–15.
- Seelbach, D. (1990). Zur Entwicklung von bilingualen Mehrwortlexica Französisch-Deutsch-Stützverbkonstruktionen und adverbiale Ausdrücke. *Lexicon und Lexikographie* 11, pp. 179–207.
- Silberztein, M. (1994). INTEX: a corpus processing system. In *Proceedings of the 15th International Conference on Computational Linguistics, COLING 94*, Kyoto, Japan, pp. 579–583.
- Villavicencio, A. (2002). Learning to distinguish PP arguments from adjuncts. In *Proceedings of the 6th Conference on Natural Language Learning, CoNLL 2002*, Taipei, Taiwan, pp. 84–90.
- Voyatzi, S. (2006). *Description morphosyntaxique et sémantique des adverbes figés en vue d'un système d'analyse automatique des textes grecs*. Thèse de doctorat en Informatique Linguistique. Université Paris -Est Marne-la-Vallée.

