A French Corpus Annotated for Multiword Nouns

Éric Laporte, Takuya Nakamura, 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, nakamura@univ-mlv.fr, voyatzi@univ-mlv.fr

Abstract

This paper presents a French corpus annotated for multiword nouns. This corpus is designed for investigation in information retrieval and extraction, as well as in deep and shallow syntactic parsing. We delimit which kind of multiword units we targeted for this annotation task; we describe the resources and methods we used for the annotation; and we briefly comment on the results. The annotated corpus is available at *http://infolingu.univ-mlv.fr/* under the LGPLLR license.

1. Introduction

Recognizing multiword nouns such as *groupes de pression* 'lobbies' in texts is useful for information retrieval and extraction because of the information that such nouns can convey. In particular, in specialized languages, most of the technical and terminological information is concentrated in multiword nouns. In addition, such recognition is likely to help resolving prepositional attachment during shallow or deep parsing: some multiword nouns contain internal prepositional phrases, and in many cases, recognising them rules out analyses where they are complements of verbs, adjectives or other nouns (Blanc *et al*., 2007). In the case of English, the same is true for the analysis of noun sequences (Vadas & Curran, 2007).

The quality of the recognition of multiword nouns depends on algorithms, but also on resources. We created a corpus of French texts annotated with multiword nouns. This corpus is freely available on the web with LGPLLR license. In this article, we survey related work, we define the target of our annotation effort, we describe the method implemented and we analyse the corpus obtained.

2. Related work

Many problems related with the notion of multiword expression (MWE) in general have been studied by linguists and lexicologists (e.g. Downing, 1977; Sag *et al*., 2001; Girju, 2005; as regards French multiword nouns: Silberztein, 1993), but textual resources annotated for MWEs are still rare and small. In the Grace corpus (Rajman *et al*., 1997), most MWEs are ignored. In the French Treebank (Abeillé *et al*., 2003), multiword nouns are annotated as such. We are not aware of other available French corpora annotated with multiword nouns. In other languages, including English, corpora annotated with MWEs are rare and small as well. In the Penn Treebank (Marcus *et al*., 1993), even such frozen nouns as *stock market* are not annotated as MWEs. Subirats & Sato (2004) report an experiment of annotating MWUs, including multiword nouns, in a Spanish corpus, and Mota *et al*. (2004) and Ranchhod (2005) in a Portuguese corpus, but the resulting annotated corpora are not publicly available. The recognition of multiword nouns is essential to identifying meaningful units in texts, and the availability of a larger corpus of annotated text is likely to shed light on the problems posed by this task.

3. Target of annotation

The target of our annotation effort is defined by the intersection of two criteria: (i) multiword expressions and (ii) nouns. In this section, we define both criteria in more detail, we define the features that we included in the annotations, and we describe the corpus. More details are provided in the guidelines which are available along with the corpus.

3.1 The multiword unit criterion

For this work, we considered a phrase composed of several words to be a multiword expression if some or all of their elements are frozen together in the sense of Gross (1986), that is, if their combination does not obey productive rules of syntactic and semantic compositionality. In the following example, *assemblée générale* ('annual general meeting', lit. 'general assembly') is a multiword noun:

(1) *Notre **assemblée générale** se tiendra vendredi*
'Our **annual general meeting** will be held on Friday'

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 restrictions such as:

* *Notre **assemblée partielle** se tiendra vendredi*
(*'Our **annual partial meeting** will be held on Friday')

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.

We annotated multiword named entities (NE) denoting places, institutions, events etc. The status of named entities with respect to compositionality is not fully consensual. We complied with the usual view that, since they follow quite specific grammatical rules, they should be considered as MWEs. However, we did not tag person names consisting of a combination of one or several first names and possibly a last name, e.g. *Gordon Brown*.

We tagged multiword nouns of functions and titles, unless they have the form *N-role de Det N-institution*, where *N-institution* is a noun denoting an institution, *Det* is a determiner, and *N-role* is a noun denoting a role assumed by a member of this institution. We consider this construction as compositional. For example, in *président de l'Assemblée nationale* 'president of the National Assembly', only *Assemblée nationale* falls in the target of our annotation task, but in *ministre de l'Économie et des Finances* 'minister of Economy and Finance', the whole phrase does.

3.2 Delimitation

The general rule to determine the delimitation of an occurrence of a multiword noun is that all and only the elements frozen with the rest of the expression should be included.

Consequently, a sequence of words should not be tagged as a multiword noun when it is included in a larger MWE. For example, the verbal idiom *aller dans le bon sens* 'be a step in the right direction' apparently includes the multiword noun *bon sens* 'good sense', but only apparently, and does not fall in the target of our annotation task. Thus, annotating multiword nouns involves analysing sentences and detecting whether frozen nominal sequences are included in larger frozen units. Such larger frozen units may be verbal idioms, as in the example above, or belong to other types, such as frozen prepositional phrases, e.g. *sur le pied de guerre* 'on a war footing', *au grand jour* 'in broad daylight; in the open', *d'un bout à l'autre de* 'from one end to the other of'. In these phrases, *pied de guerre*, *grand jour* or *bout à l'autre* should not be tagged.

When a multiword place name contains a noun denoting the type of place, we considered this noun to be a part of the multiword. For example, the nouns *océan* 'ocean' and *rue* 'street' are not included in multiword nouns when they occur in *Cet océan grisâtre l'émouvait* 'That greyish ocean moved her' or *La rue de mon enfance est de l'autre côté de l'église* 'The street of my childhood is on the other side of the church', but we analyse *océan Atlantique* and *rue de la Paix* as proper nouns.

When a multiword noun is employed with a support verb, as in *Le nouveau président donne un coup de pied dans la fourmilière*, 'The new president is kicking the anthill', the resulting construction is usually classified among multiword expressions. However, we consider that the support verb, here *donne* 'gives', is not frozen, since the noun can occur without this verb with the same meaning, as opposed to what happens with a verbal idiom. Thus, in this example, *coup de pied dans la fourmilière* 'kick in the anthill' falls in the target of our annotation task.

When a multiword noun is coordinated with another one and appears as reduced because a common part is factored, we tagged it as if it were not reduced. For example, in *accidents ferroviaires et aériens* 'rail and air crashes', the noun *accidents* 'crashes' is factored; therefore, we tagged *accidents ferroviaires* 'rail crashes' on its own, and *aériens* with the same tags as if it had the form of *accidents aériens* 'air crashes', which produces the following form[1]:

<N fs='NA:mp'>accidents ferroviaires</N> et <N fs='NA:mp'>aériens</N>

When the factored part is in the plural only because of the factoring, we tagged the multiword nouns in the singular. For example, in *les océans Atlantique et Pacifique*, both *océans Atlantique* and *Pacifique* were marked as singular.

The rules above do not apply when the whole coordination is frozen, as in *ministère de l'économie et des finances* 'ministry of Economy and Finance', which is recognizable by the impossibility to permute the coordinated parts (there is no *ministère des finances et de l'économie* 'ministry of Finance and Economy').

3.3 The noun criterion

We annotated only expressions belonging to the noun part of speech. We recognized them through the usual criteria regarding their morphosyntactic context.

Many quotations behave as nouns or names. We considered they should be tagged if they are used as titles of works. For example, the quoted sequence should be tagged in *"Autant en emporte le vent" est un film de 1939* '"Gone with the wind" is a 1939 film', but not in *Et il répondit : "Pas encore"* 'And he answered:"Not yet"'.

3.4 Features

Two types of features were included in the annotations.

(i) Each occurrence of a multiword noun was assigned a subcategory among a closed list of 13. The definition of the subcategories is based on internal morphosyntactic structure, i.e. surface constituency of the internal structure of the multiword nouns. They were described as sequences of parts of speech and syntactic categories. For example, *opinion publique* 'public opinion' is assigned a subcategory identified by the mnemonic acronym *NA*, and defined as a noun followed by an adjectival phrase. The 13 subcategories are listed in Table 1.

When a multiword noun did not strictly match any of these structures, annotators were requested to select the closest

[1] For the XML notation, see the section 3.4.

structure. For instance, *agence nationale des travailleurs d'outre-mer* 'national agency for overseas workers' is assigned the *NDN* structure, in spite of the adjective *nationale*. In case of a coordination of prepositional phrases, the multiword noun is classified as if there were only one of them: *ministre de l'emploi, de la cohésion sociale et du logement* 'minister of employment, social cohesion and housing' is assigned the *NDN* structure.

Acronym	Definition	Examples
AN	Noun with a preposed adjectival phrase or numerical determiner	*premier ministre, 35 heures*
NA	Noun with a postposed adjectival phrase	*opinion publique*
NN	Sequence of two nouns, including borrowed nouns such as *business*	*assurance-vie, pôle environnement, show-business*
VV	Sequence of two verb forms	*savoir-faire*
XV	Verb form, with a non-verb preposed modifier	*bien-être, pis-aller*
VN	Verb followed by a noun	*porte-monnaie, faire-part*
PN	Preposition followed by a noun	*après-midi, sous-traitance*
XN	Word of another category (borrowed word, prefix...) followed by a noun	*plus-value, mi-temps, stock-option*
NDN	Noun followed by a prepositional phrase with the preposition *de*	*code du travail, bien de première nécessité*
NAN	Noun followed by a prepositional phrase with the preposition *à*	*gaz à effet de serre, rappel au règlement*
NPN	Noun followed by a prepositional phrase with a preposition other than *de* or *à*	*étranger en situation irrégulière, violence contre la personne*
AAN	Noun with two preposed adjectives (coordinated or not)	*petites et moyennes entreprises*
NAA	Noun with two postposed adjectives (coordinated or not)	*produit intérieur brut, conseil économique et social*

Table 1: Morphosyntactic subcategories of multiword nouns

(ii) Inflectional features (gender and number) were also encoded in the compact form of *:ms*, *:mp*, *:fs* and *:fp*.
The syntax of the encoding follows the XML language. All features are included in an *fs* attribute, as in *<N fs='AN:ms'>Premier ministre</N>* 'prime minister'.

3.5 The corpus

The corpus we annotated comprises:
(i) the complete minutes of the sessions of the French National Assembly on October 3-4, 2006, transcribed into written style from oral French (hereafter *AS*)[2];
(ii) Jules Verne's novel *Le Tour du monde en quatre-vingts jours*, 1873 (hereafter *JV*).
Errors (e.g. *mis enoeuvre* for *mis en oeuvre* 'implemented') have not been corrected. Statistics on the corpus are displayed in Table 2.

	size (Kb)	sentences	words (tokens)	words (types)
corpus *AS*	824	5 146	98 969	18 028
corpus *JV*	1 231	3 648	69 877	19 828
whole corpus	2 055	8 794	168 846	37 856

Table 2: Size of the corpus

4. Methodology

In order to annotate the corpus, we tagged the occurrences of the multiword nouns described in a morphosyntactic lexicon, following the same method as Abeillé *et al.* (2003), Subirats & Sato (2004), Mota et al. (2004) and Ranchhod (2005); we revised the annotation manually.

4.1 The lexicon

We used the same morphosyntactic lexicon as Abeillé *et al.* (2003), so that the two corpora can be used jointly for further research. This lexicon, Delacf (Silberztein, 1990), covers the inflected forms of 100 000 lemmas. It is freely available[3] for research and business with the LGPLLR license. It is the fruit of long-term work on the basis of conventional dictionaries, corpora and introspection (Gross, 1986).

4.2 Tagging

We tagged the corpus with the Unitex platform[4] (Paumier, 2006). We used transducers in order to tag the recognized sequences with morphosyntactic features.

4.3 Manual revision

The annotation was manually validated by three experts. This validation followed guidelines, which are available along with the corpus. It involved two operations.
(i) The sequences tagged with the aid of the lexicon and Unitex were checked in order to detect cases in which the recognized sequence is in fact a part of a larger MWU. For instance, when *court terme* 'short term' occurred within the multiword adverb *à court terme* 'in the short term', the tags around *court terme* were deleted. When *ministre de l'intérieur* 'ministry of interior' occurred within the complete title *ministre de l'intérieur et de l'aménagement du territoire* 'ministry of interior and territory development', the end tag *</N>* after *intérieur* was shifted to the end of the complete title. Cases of coordinated multiword nouns (cf. section 3.2) were processed

[2] http://www.assemblee-nationale.fr/12/documents/index-rapports.asp.
[3] http://infolingu.univ-mlv.fr/english/DonneesLinguistiques/Dictionnaires/downloads.html
[4] http://igm.univ-mlv.fr/~unitex.

manually during this operation.
(ii) The text was integrally reviewed in search for multiword nouns absent from the lexicon, and thus undetected by Unitex, e.g. *passage à l'euro* 'euro changeover' or *Passe de Cheyenne* 'Cheyenne Pass'.
The experts had meetings during the annotation process in order to make it consistent. In the end, one of them reviewed the annotated corpus entirely for consistency.

5. Results

The resulting corpus is annotated with 5 054 occurrences of multiword nouns. Table 3 displays their distribution in function of the parts of the corpus and of the subcategories based on morphosyntactic structures. The percentages correspond to membership in the subcategories.

Struct.	JV corpus	JV (%)	AS corpus	AS (%)
AN	131	11.2	206	5.2
NA	206	18.7	1393	35.3
NN	267	24.2	211	5.3
VV	1	0.1	4	0.1
XV	0	0.0	4	0.1
VN	8	0.7	18	0.5
PN	11	1.0	24	0.6
XN	142	12.9	63	1.6
NDN	322	29.2	1639	41.5
NAN	7	0.6	160	4.0
NPN	6	0.5	186	4.1
AAN	1	0.1	18	0.5
NAA	1	0.1	25	0.6
Total	1103	100.0	3951	100.0

Table 3 : No. of occurrences of multiword nouns by subcategory

6. Conclusion

This paper described the annotation of a French corpus for multiword nouns. Two types of features are included in the annotations: internal morphosyntactic structure and inflectional features. This annotated corpus can be used jointly with the French Treebank (Abeillé *et al*., 2003) for research on information retrieval and extraction, automatic lexical acquisition, as well as on deep and shallow syntactic parsing.

7. Acknowledgment

This task has been partially financed by CNRS and by the Cap Digital business cluster. We thank Anne Abeillé for making the French Treebank available to us.

8. References

Abeillé, A., Clément, L., and Toussenel F. (2003). Building a Treebank for French. In A. Abeillé (Ed.), *Building and Using Parsed Corpora*, *Text, Speech and Language Technology*, 20, Kluwer, Dordrecht, pp. 165-187.

Blanc, O., Constant, M., Watrin, P. (2007). "Segmentation in super-chunks with a finite-state approach", *Proceedings of the Workshop on Finite State Methods for Natural Language Processing*, Potsdam.

Downing, P. (1977). On the Creation and Use of English Compound Nouns. *Language* 53(4), pp. 810-842.

Girju, R. et al. (2005). On the semantics of noun compounds. *Computer Speech and Language*, 19, pp. 479-496.

Gross, M. (1986). Lexicon-Grammar. The representation of compound words. In *Proceedings of the Eleventh International Conference on Computational Linguistics*, Bonn, West Germany, pp. 1-6.

Marcus, M., Santorini, B., Marcinkiewicz, M.A. (1993). "Building a large annotated corpus of English: the Penn Treebank", *Computational Linguistics* 19(2), pp. 313-330.

Merlo, P. (2003). Generalised PP-attachment Disambiguation using Corpus-based Linguistic Diagnostics. In *Proceedings of the Tenth Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, pp. 251-258.

Mota, C. Carvalho, P. Ranchhod, E. (2004). Multiword Lexical Acquisition and Dictionary Formalization. In Michael Zock (ed.), *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*, COLING, Geneva, pp. 73-76.

Paumier, S. (2006). *Unitex Manual*. Université Paris-Est. http://igm.univ-mlv.fr/~unitex/manuel.html.

Rajman, M., Lecomte, J., Paroubek, P. (1997). Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique. Rapport GRACE GTR-3-2.1.

Ranchhod, E. (2005). "Using Corpora to Increase Portuguese MWE Dictionaries. Tagging MWE in a Portuguese Corpus". In: *Proceedings from The Corpus Linguistics Conference Series, Vol. 1, no. 1, Corpus Linguistics 2005*.

Sag, I., Baldwin, T., Bond, F., Copestake, A., Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In Gelbuk (ed.), *Computational Linguistics and Intelligent Text Processing: Third International Conference CICLing 2002*, Springer-Verlag, Heidelberg/Berlin, pp. 1--15.

Silberztein, M. (1990). "Le dictionnaire électronique des mots composés". *Langue Française* 87, pp. 71-83, Paris : Larousse.

Silberztein, M. (1993). "Les groupes nominaux productifs et les noms composés lexicalisés", *Lingvisticae Investigationes* 17:2, Amsterdam/Philadelphia: Benjamins, pp. 405-426.

Subirats, C., Sato, H. (2004). Spanish FrameNet and FrameSQL. *4th International Conference on Language Resources and Evaluation. Workshop on Building Lexical Resources from Semantically Annotated Corpora, Lisbon (Portugal), May 2004*.

Vadas, D., Curran, J.R. (2007). " Adding Noun Phrase Structure to the Penn Treebank", In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 240–24.