A French Corpus Annotated for Multiword Expressions with Adverbial Function

Eric Laporte, Takuya Nakamura, Stavroula Voyatzí

To cite this version:


HAL Id: halshs-00286541
https://halshs.archives-ouvertes.fr/halshs-00286541v2
Submitted on 11 Jun 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A French Corpus Annotated for Multiword Expressions with Adverbial Function

Eric Laporte, Takuya Nakamura, Stavroula Voyatzi

Université Paris-Est
Institut Gaspard-Monge - 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 expressions (MWEs) with adverbial function. This corpus is designed for investigation on information retrieval and extraction, as well as on deep and shallow syntactic parsing. We delimit which kind of MWEs we annotated, we describe the resources and methods we used for the annotation, and we briefly comment the results. The annotated corpus is available at http://infolingu.univ-mlv.fr/ under the LGPLLR license.

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 such adverbials can 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 analyses where they are arguments or noun modifiers.

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

2. Related work

Corpora annotated with multiword adverbs are rare and small. In the Grace corpus (Rajman et al., 1997), most multiword units are ignored. In the French Treebank (Abeillé et al., 2003), 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). We are not aware of other available French corpora annotated with multiword adverbs. In other languages, including English, corpora annotated with multiword units are rare and small as well.

3. Target of annotation

The target of our annotation effort is defined by the intersection of two criteria: (i) multiword expressions and (ii) adverbial function. 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.

3.1 Multiword expression 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, de nos jours ‘nowadays’, lit. ‘of our days’) is a multiword adverb:

(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 (Freckleton, 1985; Machonis, 1985; Silberztein, 1993; Lamiroy, 2003): 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

---

1 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 and lies in complex linguistic notions (Villavicencio, 2002; Merlo, 2003). The task is therefore felt as too difficult by most researchers in language processing, whose main background is in information technology. However, the distinction in question is essential to identifying the semantic core of a sentence, and the availability of a larger corpus of annotated text is likely to shed light on the problems posed by this task.

2 That can be empirically checked only after a lexicon and a grammar for the same language are complete and compatible.
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:

* Il est facile de nos semaines de s’informer
* It is easy to get informed now

Multiword expressions include many different subtypes,
varying from entirely fixed expressions to syntactically
more flexible expressions (Sag et al., 2002). We
annotated expressions undergoing variations. 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

‘With (his + *my) own hands, he built a house’

3.2 Adverbial function

We annotated only expressions with adverbial function,
or circumstantial complements, i.e. complements which
are not objects of the predicate of the clause in which
they appear. We recognised them through criteria (Gross
1986, 1990a, 1990b) 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: derived (de...’), 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
named entities of time (lundi 20 on Monday 20’). We
annotated NEs only when they have an adverbial function,
as in: Jean arrive lundi 20 ‘John arrives on
Monday 20’. NEs of other categories, such as places,
persons, events, etc., are usually not adverbials.

3.3 Features

Two types of features were included in the annotations.
(i) Each occurrence of a multiword adverb was assigned
one internal morphosyntactic structure or semantic type
among 19. 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
categorical attributes. For example, à la dernière minute ‘at
‘nightfall’ is assigned a structure identified by the
mnemonic acronym MPA, 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.
For named entities, this feature encodes the semantic
type: date, duration, time or frequency, in conformity
with the typology of the Infom@gic project (Martineau
et al., 2007). The 19 structures and semantic types are
listed in Table 1. In this table, N stands for a free noun
phrase, and W for a variable ranging over verb
complements. Other symbols are easy to interpret: Prép,
Dét, Adj, V, Conj...

3.4 The corpus

The corpus we annotated includes: (a) 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) and (b) Jules Verne’s novel Le
Tour du monde en quatre-vingts jours, 1873 (hereafter
JV). Errors (e.g. mis en oeuvre for mis en œuvre ‘implemented’) have not been corrected. Statistics on
the corpus are displayed in Table 2.

---

Table 1: Morphosyntactic structures and semantic types

<table>
<thead>
<tr>
<th>Identifiers</th>
<th>Structures</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
<td>Prép C</td>
<td>en bref</td>
</tr>
<tr>
<td>PDTEC</td>
<td>Prép Dét C</td>
<td>de nos pures</td>
</tr>
<tr>
<td>PAC</td>
<td>Prép Adj C</td>
<td>à la dernière minute</td>
</tr>
<tr>
<td>PCA</td>
<td>Prép C Adj</td>
<td>à la nuit tombante</td>
</tr>
<tr>
<td>PCDC</td>
<td>Prép C de C</td>
<td>dans le minute du possible</td>
</tr>
<tr>
<td>PPOC</td>
<td>Prép C Préc C</td>
<td>dois pieds à la tête</td>
</tr>
<tr>
<td>PCONJ</td>
<td>Prép C Conj C</td>
<td>en avant et pour tant</td>
</tr>
<tr>
<td>PCDN</td>
<td>Prép C de N</td>
<td>au moyen de N</td>
</tr>
<tr>
<td>PCPN</td>
<td>Prép C Préc N</td>
<td>par rapport à N</td>
</tr>
<tr>
<td>PV</td>
<td>Prép V W</td>
<td>à dire vrai</td>
</tr>
<tr>
<td>PF</td>
<td>P (frozen clause)</td>
<td>jusqu’à ce que mort s’ensuive</td>
</tr>
<tr>
<td>PECO</td>
<td>(Adj) comme C</td>
<td>comme ses pieds</td>
</tr>
<tr>
<td>PVCO</td>
<td>(V. comme C)</td>
<td>comme un cheveau sur la onde</td>
</tr>
<tr>
<td>PPCO</td>
<td>(V. comme Prép C)</td>
<td>comme dans du beurre</td>
</tr>
<tr>
<td>PJJC</td>
<td>Conj C</td>
<td>mais enfin et surtout</td>
</tr>
<tr>
<td>DATE</td>
<td>Named Entities</td>
<td>le 22 mai 2003</td>
</tr>
<tr>
<td>DURATION</td>
<td>Named Entities</td>
<td>pendant vingt-quatre heures</td>
</tr>
<tr>
<td>TIME</td>
<td>Named Entities</td>
<td>à huit heures du soir</td>
</tr>
<tr>
<td>FREQUENCY</td>
<td>Named Entities</td>
<td>deux fois par jour</td>
</tr>
</tbody>
</table>

Notes:
5 The point is that this blocking of distributional variation (and
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.
6 We annotated phrases which comprise a frozen part and a free
part, e.g. au moyen de ce bouton ‘with the aid of this switch’, in
which au moyen de ‘with the aid of’ is frozen, and ce bouton
‘this switch’ is a distributionally free noun phrase embedded in
the global phrase. In such cases, we delimited the embedded
free part with tags (cf. section 4.2). Finally, we annotated
named entities (NEs) of date and duration. The status of named
entities with respect to compositionality is not fully consensual:
however, we complied with the usual view that, since they
follow quite specific grammatical rules, they should be
considered as multiword expressions.
In order to annotate the corpus, we tagged the occurrences of the expressions described in a syntactic-semantic lexicon of adverbials, as Abeillé et al. (2003), Baptista (2003) for Portuguese, and Català & Baptista (2007) for Spanish; we tagged NEs of date, duration, time, and frequency through a set of local grammars, as Friburger & Maurel (2004); then, we revised the annotation manually.

### 4.1 The lexicon

We used the same syntactic-semantic lexicon (Gross, 1990a) as Abeillé et al. (2003), so that the two corpora can be used jointly for further research. This lexicon has 6,800 entries. It is freely available\(^5\) for research and business under the LGPL license. It was constructed on the basis of conventional dictionaries, grammars, corpora and introspection, within the Lexicon-Grammar methodology (Gross, 1986; 1994). It takes the form of a set of Lexicon-Grammar tables such that of Table 3, which displays a sample of the lexical items with the PCA morphosyntactic structure.

![Table 3: Sample of the table of entries with the PCA morphosyntactic structure](image)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>2</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>3</td>
<td>agr</td>
<td>dans</td>
<td>les</td>
<td>de</td>
<td>les</td>
<td>plus</td>
<td>breifs</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>4</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>5</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>+</td>
<td>se produire</td>
<td>à</td>
<td>cette</td>
<td>écoupe</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>


- 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. The features provided by the lexicon were used to annotate the occurrences.

### 4.2 Tagging

We tagged the corpus with the Unitex system (Paumier, 2006). Many multiword adverbs are entirely fixed expressions, but others present variations, such as grammatical agreement (cf. example (2), section 3.1), permutations and omissions. Due to these variations, we tagged them with finite-state transducers (FST): the input part of these transducers recognises the expressions and their variants, and the output part inserts the tags. Like Català & Baptista (2007), we used lexicalised transducers, i.e. one for each lexical item, and we generated them with the technique of parameterised graphs (Roche, 1999) modified by Silberztein (1999).

Multiword adverbs with a free prepositional phrase modifier (morphosyntactic structures PCPN and PCPN) were annotated semi-automatically as follows (‘NP’ if the free complement is occupied by a noun phrase, ‘S’ if it is occupied by a clause):

1. \(<ADV \ fs=\ "PCDN"\> compte tenu de \(<NP\> vos\ ambitions</NP></ADV>\)
   - ‘taking into account your ambitions’
2. \(<ADV \ fs=\ "PCDN"\> compte tenu de \(<S\> ce que tout va bien</S></ADV>\)
   - ‘taking into account that everything is OK’

Named entities with temporal value (cf. section 3.2) were automatically tagged by using FST methods similar to those applied for multiword adverbs.

### 4.4 Manual revision

The annotation was manually reviewed by three experts. This validation followed guidelines, which are available along with the corpus. It involved two operations.

1. The sequences tagged with the aid of the lexicon and Unitex were checked in order to detect cases when the recognised sequence is in fact a part of a larger MWE.
   - (i) The text was integrally reviewed in search for 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).
   - (ii) The text was integrally reviewed in search for 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).

2. When the embedded free part of a multiword adverb is a coordination, we tagged it manually:
   - \(<ADV \ fs=\ "PCDN"\> en terms de \(<NP\> santé</NP> et \ d'\(<NP\> éducation</NP></ADV>\)
   - ‘in terms of health and education’

   The text was integrally reviewed in search for multiword adverbs absent from the lexicon, and thus undetected by Unitex, e.g. de plus ‘moreover’ or for le moins ‘at least’. This required for the annotators to identify the syntactic structure of each sentence in the corpus. We had meetings during the annotation process in order to make it consistent.

Table 2: Size of the corpus

<table>
<thead>
<tr>
<th>corpus</th>
<th>size (Kb)</th>
<th>sentences</th>
<th>tokens</th>
<th>types</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS</td>
<td>824</td>
<td>5 146</td>
<td>98 969</td>
<td>18 028</td>
</tr>
<tr>
<td>JV</td>
<td>1 231</td>
<td>3 648</td>
<td>69 877</td>
<td>19 828</td>
</tr>
<tr>
<td>total</td>
<td>2 055</td>
<td>8 794</td>
<td>168 846</td>
<td>37 856</td>
</tr>
</tbody>
</table>
5. Results

This corpus is annotated with 4,383 occurrences of MWEs with adverbial function. 1,118 of them correspond to multiword adverbials with conjunctive function in discourse. They represent about 3.5% of the overall of simple word occurrences occurring in the whole corpus. Table 4, below, shows the number of occurrences of annotated MWEs. The lines of the table correspond to the morphosyntactic structures and semantic types.

<table>
<thead>
<tr>
<th>Identifiers</th>
<th>JV corpus</th>
<th>JV % cover</th>
<th>JV Adv Conj</th>
<th>AS corpus</th>
<th>AS % cover</th>
<th>AS Adv Conj</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
<td>333</td>
<td>0.70</td>
<td>150</td>
<td>453</td>
<td>0.70</td>
<td>274</td>
</tr>
<tr>
<td>PSC/PSTC</td>
<td>254</td>
<td>0.93</td>
<td>74</td>
<td>235</td>
<td>0.34</td>
<td>92</td>
</tr>
<tr>
<td>PAC</td>
<td>87</td>
<td>0.15</td>
<td>22</td>
<td>146</td>
<td>0.23</td>
<td>76</td>
</tr>
<tr>
<td>PCA</td>
<td>81</td>
<td>0.15</td>
<td>18</td>
<td>87</td>
<td>0.19</td>
<td>24</td>
</tr>
<tr>
<td>PCC</td>
<td>90</td>
<td>0.07</td>
<td>9</td>
<td>39</td>
<td>0.09</td>
<td>11</td>
</tr>
<tr>
<td>PCPC</td>
<td>56</td>
<td>0.05</td>
<td>3</td>
<td>58</td>
<td>0.07</td>
<td>0</td>
</tr>
<tr>
<td>PCONJ</td>
<td>75</td>
<td>0.03</td>
<td>1</td>
<td>21</td>
<td>0.05</td>
<td>1</td>
</tr>
<tr>
<td>PCDF</td>
<td>246</td>
<td>0.03</td>
<td>14</td>
<td>252</td>
<td>0.10</td>
<td>84</td>
</tr>
<tr>
<td>PPN</td>
<td>88</td>
<td>0.20</td>
<td>47</td>
<td>117</td>
<td>0.15</td>
<td>26</td>
</tr>
<tr>
<td>PF</td>
<td>56</td>
<td>0.07</td>
<td>5</td>
<td>74</td>
<td>0.10</td>
<td>23</td>
</tr>
<tr>
<td>PEOCO</td>
<td>15</td>
<td>0.14</td>
<td>11</td>
<td>27</td>
<td>0.05</td>
<td>21</td>
</tr>
<tr>
<td>SPCO</td>
<td>1.00</td>
<td>0.00</td>
<td>0</td>
<td>1.00</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>PVOCO</td>
<td>3.00</td>
<td>0.01</td>
<td>0</td>
<td>3.00</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>PCOCO</td>
<td>2.00</td>
<td>0.01</td>
<td>0</td>
<td>2.00</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td>PUCO</td>
<td>2.00</td>
<td>0.01</td>
<td>1</td>
<td>2.00</td>
<td>0.01</td>
<td>3</td>
</tr>
<tr>
<td>DATE</td>
<td>243</td>
<td>0.49</td>
<td>18</td>
<td>262</td>
<td>0.42</td>
<td>47</td>
</tr>
<tr>
<td>DURATION</td>
<td>138</td>
<td>0.22</td>
<td>15</td>
<td>118</td>
<td>0.13</td>
<td>3</td>
</tr>
<tr>
<td>TIME</td>
<td>96</td>
<td>0.18</td>
<td>0</td>
<td>20</td>
<td>0.02</td>
<td>0</td>
</tr>
<tr>
<td>FREQUENCY</td>
<td>32</td>
<td>0.00</td>
<td>7</td>
<td>40</td>
<td>0.05</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>1,769</td>
<td>3.28</td>
<td>471</td>
<td>2,564</td>
<td>3.48</td>
<td>167</td>
</tr>
</tbody>
</table>

Table 4: MWEs with adverbial function in the corpus

6. Conclusion

This paper described the design of a French corpus annotated for MWEs with adverbial function. Various types of features are included in the annotations: the morphosyntactic structure, special functions in discourse (e.g. the conjunctive function) and the semantic types of named entities of time. This annotated corpus can be used jointly with the French Treebank (Abéillé 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


