1-1-2003

The TALANA treebank for French

Anne Abeillé
Lionel Clement
Alexandra Kinyon

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/pwpl/vol8/iss1/2
For more information, please contact libraryrepository@pobox.upenn.edu.
The TALANA treebank for French
The TALANA Treebank for French*

Anne Abeillé, Lionel Clément and Alexandra Kinyon

1 Abstract

This paper presents the first linguistic results exploiting the new annotated corpus for French developed at Talana-Paris 7 (Abeillé et al. 2000). The corpus comprises one million words fully annotated and disambiguated for parts of speech, inflectional morphology, compounds and lemmas, and partially annotated with syntactic constituents. It is representative of contemporary normalized written French, and covers a variety of authors and subjects (economy, literature, politics, etc.), with extracts from newspapers ranging from 1989 to 1993.

After explaining how this corpus was built, we present some linguistic results obtained when searching the corpus for lexical or syntactic frequencies, for lexical or syntactic preferences, and explain why we think some of these results are relevant both for theoretical linguistics and psycholinguistics.

2 Building a Treebank for French

Similarly to the Penn Treebank (Marcus et al. 1993), we distinguish a tagging and a parsing phase, and define a process of automatic annotation followed by a systematic manual validation and correction. Similarly to the Suzanne Corpus (Sampson 1995) or the Prague treebank (Hajicova et al. 1998), we rely on several types of morphosyntactic and syntactic annotations for which we define extensive guidelines. Our goal is to provide a theory neutral, surface-oriented, error-free treebank for French.

The corpus is made of extracts from the newspaper LeMonde, made publicly available for research purposes through LDC. It comprises roughly one million words. With compounds amalgamated and not counting punctuation marks, it comprises 870,000 tokens, using 17,000 different lemmas, making up about 32,000 independent sentences.

*This project was sponsored by Institut Universitaire de France (IUF), by CNRS (as part of the CLIF project), by LORIA (as part of the CALIN project) and by Au-pelf-Urf (as part of the Corfrans Project).

2.1 Morpho-syntactic Annotation

Our corpus has been annotated and fully disambiguated for morphosyntactic annotation (with longitudinal human validation and double checks).\(^1\) We have a richer morphosyntactic tagset than most annotated corpora (218 different tags, which are valid combinations of the notations presented in appendix 1).

We define a complete morphosyntactic tag as follows:

1. Part of Speech (POS); e.g. *Determiner*.
2. Subcategorization; e.g. *possessive or cardinal*.
3. Inflection; e.g. *masculine singular*.
4. Lemma (canonical form).
5. Parts (with similar morphosyntactic tags) for compounds.

For parts of speech, we made traditional choices, except for weak pronouns that were given a POS of their own (Clitic) according to the generative linguistic tradition, and foreign words which receive a special POS (ET). Punctuation marks are divided between strong (clause markers) and weak (all the others). Most typographical signs (including ‘%’, numbers and abbreviations) are assigned a traditional POS (usually Common Noun).

Because of the rich morphology of French, we chose to annotate more than just parts of speech. In order to allow for multiple views on the corpus, we annotated both compounds and parts of compounds with the same tagset, so a user can choose to retain or to ignore our choices for compounds.

Difficult cases involved tagging numbers, tagging weak pronouns (clitics), choosing between adjective and past participle, between proper and common Noun (for unknown words), between Prep and (indefinite or partitive) Det (for *de*). For numbers, we depart from Multex guidelines in choosing the same tagset as for other words. The annotators thus had to choose between:

- determiner: *Deux hommes sont venus* (Two men came)
- pronoun: *Il en a accueilli deux* (He welcomed two of them)
- adjective: *Les deux hommes sont venus* (The two men came)
- noun: *Le joueur a misé sur le deux* (The player bet on the two)

For clitic pronouns, we simplified the usual case system and kept only nominative, objective, and reflexive subcategories, since assigning the right

\(^1\)We used a tagger (Reyes 1998) based on Brill’s rule-based POS tagger and developed especially for this purpose.
case (or no case at all for uses as inherent clitics or mediopassive) is part of syntactic analysis and will be done (partly automatically) in the second phase of the project. Another difficulty is that most clitic forms in French are ambiguous with respect to gender (je, leur, les) or number (se) or both (y, en). The annotator thus had to find their antecedent to properly annotate their morphosyntax.

Most difficult cases involved ambiguous grammatical words (such as tous ‘all’ or que ‘that’), the tagging of which is a matter of debate among linguists since it depends on the syntactic analysis of notoriously complex constructions (cleft sentences, comparatives etc). In such cases, we made obviously debatable choices: our main goals were to be explicit (in the documentation), consistent (throughout the corpus) and theory neutral (so that our tagging is compatible with several syntactic analyses).

2.2 Syntactic Annotation

Contrary to tagging, precise language specific guidelines are usually missing for syntactic annotation. In order to provide annotations reusable by researchers from various backgrounds, we chose to annotate both constituency and functional relations. We focus here on constituency annotations.

We chose surface and shallow annotations, compatible with various syntactic frameworks, and easily learnable for human annotators.

The following information will be contained in each syntactic tag:

1. Main category (e.g. S, PP, NP...)
2. Eventual subcategory (e.g. Rel for relative clauses)
3. Surface function (e.g. Subj, Object for NPs)
4. Opening or closing boundaries (<?, </>)
5. Valence (e.g. ditransitive) for verbal nuclei

For the moment, we only have annotated phrasal names (category and subcategory) and phrasal (i.e. constituent) boundaries, using a robust rule-based shallow parser described in (Kinyon 2001), (Clément and Kinyon 2000). This automatic bracketing is followed by a phase of systematic and longitudinal human checking and correction, using an Emacs-based tool (with graphical display) especially designed by Michel Simard and Lionel Clément. The task of the annotator is to check both constituency names and phrase boundaries, especially for PPs left unattached by the shallow-parser.

We chose to only annotate major phrases, with little internal structure (we have determiners and modifying adjectives at the same level in the noun phrase for example). For the sake of simplicity, we make a parsimonious use of unary phrases. For rigid sequences of categories, such as dates or ad-
dresses, it is difficult to determine the head, and we have one global NP with no internal constituents.

We do not have discontinuous nor empty constituents, since the corresponding information (such as passive or missing subject) will be encoded directly at the functional level.

We use 12 different tags for constituents (see appendix 1). We made two specific choices, regarding verbal phrases, and regarding coordinated phrases, in accordance with the specificity of French.

For verbal phrases, we only annotate the minimal verbal nucleus (clitics, auxiliaries, negation and verb), because the traditional VP (with complements) is subject to much linguistic debate and is often discontinuous in French. For coordination, we do not necessarily embed conjuncts inside a coordinating phrase, in order to be able to cope with non constituent coordination and coordination of unlike constituents. We consider the first conjunct as the head and annotate each following conjunct with a specific category COORD.

Most of the difficult cases were with PP attachment, or scope of coordination, for which a deep understanding of the sentences is necessary. The only remaining ambiguities are thus only spurious ones (with the same interpretation) and we chose to get rid of them by the Attach high heuristics.

3 Exploiting the Annotated Corpus

There are a large number of uses that can be made of this annotated corpus. We present here some results regarding lexical or syntactic frequency and lexical or syntactic preferences, which are of relevance both for psycholinguistics and for computational linguistics.

3.1 Lexical Frequency

Lexical frequencies for French have usually been computed on raw data (Catach, Julliand). As shown for example by (Silberztein 1993), such counts are necessarily erroneous given the high proportion of ambiguous forms.

Let us see how the part of speech disambiguation performed on our corpus improves such calculations.

If we rank the forms by frequency, we obtain the list in the second column (table 1) as the most common forms, which only comprises function words (prepositions, determiners, conjunctions) and is comparable with what other authors find on different French corpora. But most of these forms are in fact ambiguous: de can be a preposition or a determiner, le can be a determiner or a pronoun, en can be a preposition (in) or a clitic pronoun (of it). If one is interested in the most common words in the corpus, it is thus necessary
on one hand to discriminate these ambiguous forms and on the other hand to
gather different inflections of the same word (d' and de for the preposition
DE, le, la, les, l' for the determiner LE, etc).

If we do this and rank the forms by (disambiguated) lemma, we obtain
the list in the third column which is quite different. Now the most common
word is the determiner LE and some verbs (être, avoir) are among the 10
most frequent words.

<table>
<thead>
<tr>
<th>Lexical frequency</th>
<th>by form</th>
<th>by lemma+POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>de (Prep or Det)</td>
<td>LE (le, la, les, l') Det</td>
</tr>
<tr>
<td>2nd</td>
<td>le (Det or CL)</td>
<td>de (de, d') Prep</td>
</tr>
<tr>
<td>3rd</td>
<td>les (Det or CL)</td>
<td>à Prep</td>
</tr>
<tr>
<td>4th</td>
<td>la (Det or CL)</td>
<td>un (un, une, des, de, d') Det</td>
</tr>
<tr>
<td>5th</td>
<td>à</td>
<td>être (suis, est etc) V</td>
</tr>
<tr>
<td>6th</td>
<td>l' (Det or CL)</td>
<td>et CC</td>
</tr>
<tr>
<td>7th</td>
<td>et</td>
<td>avoir (ai, a etc) V</td>
</tr>
<tr>
<td>8th</td>
<td>en (Prep or CL)</td>
<td>il (il, ils, elle, elles) CL</td>
</tr>
<tr>
<td>9th</td>
<td>un</td>
<td>en (Prep)</td>
</tr>
</tbody>
</table>

Table 1. Lexical frequencies

If we now rank the categories themselves, we obtain the figures in table
2, which are again quite different. Contrary to what highly frequent forms
show, the most common lexical categories are not function words such as
determiners, pronouns, prepositions or auxiliary verbs.

<table>
<thead>
<tr>
<th>POS</th>
<th># occurrences</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns</td>
<td>226879</td>
<td>24.5%</td>
</tr>
<tr>
<td>Determiners</td>
<td>156008</td>
<td>16.8%</td>
</tr>
<tr>
<td>Prepositions</td>
<td>134753</td>
<td>14.6%</td>
</tr>
<tr>
<td>Punctuation marks</td>
<td>122448</td>
<td>13%</td>
</tr>
<tr>
<td>Verbs</td>
<td>105901</td>
<td>11.4%</td>
</tr>
<tr>
<td>Adjectives</td>
<td>60310</td>
<td>6.5%</td>
</tr>
<tr>
<td>Adverbs</td>
<td>45204</td>
<td>4%</td>
</tr>
<tr>
<td>Conjunctions</td>
<td>30623</td>
<td>3.3%</td>
</tr>
<tr>
<td>Clitics</td>
<td>26055</td>
<td>2.8%</td>
</tr>
<tr>
<td>Pronouns (other)</td>
<td>17172</td>
<td>1.8%</td>
</tr>
</tbody>
</table>

Table 2. Repartition of POS in the tagged corpus

---

2This repartition may be specific to newspaper genre in French.
Obviously more fine-grained calculations are called for (for example regarding the different types of adverbs or adjectives). If one considers the relative frequency of functionally marked forms, namely relative and clitic pronouns, one gets the following results (on the whole corpus):

**Relative pronouns:**
- subject (qui without prep) 6291 (61%)
- direct object (que, qu') 1565 (15.2%)
- genitive (dont) 1076 (10.4%)
- locative (ou) 782 (7.6%)
- indirect object (prep+qui, quoi, lequel) 539 (5.2%)
- others 0.3%

This repartition is reminiscent with what was found by Keenan and Hawkins (1987) on English newspaper texts, which confirms Keenan and Comrie's universal relative accessibility hierarchy:

- subject relative (who, that) 46%
- direct object relative (whom, that) 24%
- indirect object relative (prep whom/which) 15%
- genitive relative (whose) 5%
- others (locative...) 10%

In French, the preference for subject relatives is much stronger than for English, and the relative frequency of the genitive (dont) is also higher, maybe due to the frequent use of dont relative clauses with a resumptive pronoun in French newspapers (*Un problème dont on sait qu'il est difficile, 'a problem which one knows it is difficult').

We also check that the same functional hierarchy is also observed for clitic pronouns (which are the other type of non-canonical realization in French). The observed relative frequency is the following (using our simplified marking which does not distinguish direct from indirect objects):

**Clitics (weak personal pronouns)**
- CL subject (je, tu, il(s), elle(s), ce) 14243 (54.8%)
- CL reflexive (me, te, se, nous, vous) 6567 (25.3%)
- CL object (me, te, le, la les, lui, leur, nous, vous) 3124 (12.2%)
- CL oblique (en, y) 2018 (7.6%)

Again, the subject pronouns are much more frequent than the other ones, and this shows a strong correlation between non-canonical realization and functional accessibility.
3.2 Lexical Preferences

When one considers syntactically ambiguous forms, it is usually the case that the probabilities of the different parts of speech are quite unequal (cf. Church 1988), and this is why stochastic taggers perform reasonably well (with a small tagset).

Psycholinguists also claim that syntactic preferences can be associated to lexical items, but it is difficult to claim that a specific preference has to be learned for each ambiguous form. This is why we have looked for more general preference principles, that can be helpful for developing automatic POS taggers but also that can shed light on human parsing strategies.

At the tokenization (or word split) level, we first checked the well known preference for compounds. We took the sequences which are possibly ambiguous between compounds and non-compound sequences and compute their respective number of occurrences. Examples of such pairs would be:

EN FAIT: compound Adv (in fact) OR en:Clitic fait:Verb (makes it)
D'AILLEURS: compound Adv (besides) OR d':Prep + ailleurs:N (from elsewhere)

Some results are shown table 3.

<table>
<thead>
<tr>
<th>Possible compound</th>
<th># occurrences as compound</th>
<th># occurrences as non-compound</th>
</tr>
</thead>
<tbody>
<tr>
<td>pomme de terre</td>
<td>100 % (NC)</td>
<td>0 % NC Prep NC</td>
</tr>
<tr>
<td>D'abord</td>
<td>154 (97 %) Adv</td>
<td>5 (3%) Prep NC</td>
</tr>
<tr>
<td>alors que</td>
<td>231 (96%): CS</td>
<td>8 (4%) Adv CS</td>
</tr>
<tr>
<td>plus de</td>
<td>305 (60%) Prep</td>
<td>(40%) Adv Prep or Det</td>
</tr>
<tr>
<td>le plus</td>
<td>123 (39%) Adv</td>
<td>(61%) Det Adv</td>
</tr>
<tr>
<td>sur ce</td>
<td>0 (Adv)</td>
<td>65 (100%) Prep Det</td>
</tr>
</tbody>
</table>

Table 3. Respective proportion of compound and non-compound categories.

The preference is attested (more than 93% of occurrences as a compound on average) but depends on the categories involved. For nominal and verbal compounds (usually made of Nouns, Verbs and Adjectives) the compound interpretation covers almost 100% of the occurrences. For adverbial compounds, the preference is lower, and there are exceptions such as ‘sur ce’ or ‘le plus’ in table 3. This lower preference can be explained by an overriding preference for the grammatical categories (Clitic, Determiner, Preposi-
tion... see below) associated with the words involved in the non-compound interpretation.

We check that the preference for the compound interpretation is a lexical preference because the total number of occurrences of compounds in the corpus is much lower than that of non-compound words (50614=6.2% vs 765953=93.8 % ignoring punctuation).

At the tagging (or POS disambiguation) level, we found a strong lexical preference for grammatical versus lexical categories. We took grammatical categories as closed class of function words (Determiners, Prepositions, Clitics and other Pronouns, Subordinating and coordinating conjunctions) whereas lexical ones are the open class ones (V, Adj, N, Adv).

We took the lexical forms ambiguous between these two classes and computed the respective frequency of their occurrences in the corpus. Examples of such pairs are:

CAR: car:conjunction (since) OR car:n (bus)
OUTRE: outre:prep (in addition of) OR outre:n (drinking container)
ENTRE: entre:prep (between) OR entre:v (enter)

Some results are shown table 4:

<table>
<thead>
<tr>
<th>Ambiguous form</th>
<th>Total # occurrences</th>
<th>Occurrences with lexical category</th>
<th>Occurrences with grammatical category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car</td>
<td>235</td>
<td>5 (2.1%) noun</td>
<td>230 (97.8%) C conj</td>
</tr>
<tr>
<td>Cela</td>
<td>284</td>
<td>1 (0.3%) verb</td>
<td>283 (99.7%) pronoun</td>
</tr>
<tr>
<td>Dans</td>
<td>5341</td>
<td>0 (0%) noun</td>
<td>5341 (100%) preposition</td>
</tr>
<tr>
<td>devant</td>
<td>285</td>
<td>33 (11.5%) verb</td>
<td>252 (88.4%) preposition</td>
</tr>
<tr>
<td>Entre</td>
<td>1195</td>
<td>23 (1.9%) verb</td>
<td>1172 (98%) preposition</td>
</tr>
<tr>
<td>envers</td>
<td>25</td>
<td>3 (12%) noun</td>
<td>22 (88%) preposition</td>
</tr>
<tr>
<td>La</td>
<td>24471</td>
<td>1 (0%) noun</td>
<td>24470 (100%) det, clitic</td>
</tr>
<tr>
<td>Lui</td>
<td>763</td>
<td>0 (0%) verb (luiire)</td>
<td>763 (100%) clitic, pronoun</td>
</tr>
<tr>
<td>Or</td>
<td>189</td>
<td>30 (15.9%) noun</td>
<td>159 (84.1%) C coord</td>
</tr>
<tr>
<td>Si</td>
<td>989</td>
<td>0 (0%) noun</td>
<td>989 (100%) C sub, Adv</td>
</tr>
<tr>
<td>Son</td>
<td>2427</td>
<td>2417 (99.6%) noun</td>
<td>10 (0.4%) det</td>
</tr>
<tr>
<td>Sous</td>
<td>359</td>
<td>25 (7%) noun</td>
<td>334 (93%) prep</td>
</tr>
<tr>
<td>Ton</td>
<td>31</td>
<td>22 (71%) noun</td>
<td>9 (29%) det</td>
</tr>
</tbody>
</table>

Table 4. Relative frequencies of lexical vs grammatical categories for ambiguous forms.
Overall, we found an overwhelming proportion of uses as grammatical categories (more than 95% on the average, sometimes 100%).

Again, we check that this is a lexical preference because the total number of occurrences of grammatical categories is not higher than that of lexical categories in the corpus as a whole (43.6% vs 46.4%), as shown in table 2 above.

4 Conclusion and Future Work

We have presented a syntactically annotated corpus for French, fully disambiguated and manually validated, and some preliminary investigations. Some of these investigations have confirmed well known frequencies or lexical preferences, others have brought to light new frequencies and new preferences that should be confirmed on other corpora.

Future inquiries on this corpus comprise attachment preferences, especially for relative clauses or PPs following two candidate head Nouns, in collaboration with psycholinguists. Comparisons will also have to be made with other treebanks for other languages.

Future annotation involves assigning a grammatical function to each major phrase. This will permit more investigations, for example on subject inversion.

The corpus is distributed as a linguistic resource and is already being used by a few teams in France and elsewhere.
Appendix 1 Tagset of the TALANA corpus

<table>
<thead>
<tr>
<th>POS</th>
<th>Subcategorization</th>
<th>Morphology</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>Common, proper</td>
<td>f,m + s,p</td>
<td>Nouns</td>
</tr>
<tr>
<td>A</td>
<td>Cardinal, ordinal, possessive, qualifier, indefinite, interrogative</td>
<td>f,m + s,p + 1,2,3</td>
<td>Adjectives</td>
</tr>
<tr>
<td>Adv</td>
<td>-, inter, exclam, negative</td>
<td>-</td>
<td>Adverbs</td>
</tr>
<tr>
<td>P</td>
<td>-</td>
<td>-</td>
<td>Prepositions</td>
</tr>
<tr>
<td>D</td>
<td>Card, dem, def, indef, exclam, negative, poss, inter, partitive</td>
<td>f,m + s,p + 1,2,3</td>
<td>Determiners</td>
</tr>
<tr>
<td>CL</td>
<td>subj, refl, obj, -</td>
<td>f,m + s,p + 1,2,3</td>
<td>Clitic pronouns</td>
</tr>
<tr>
<td>PRO</td>
<td>Inter, pers, negative, poss, rel, indef</td>
<td>f,m + s,p + 1,2,3</td>
<td>Other pronouns</td>
</tr>
<tr>
<td>C</td>
<td>Subord, Coord</td>
<td>-</td>
<td>Conjunctions</td>
</tr>
<tr>
<td>I</td>
<td>-</td>
<td>-</td>
<td>Interjections</td>
</tr>
<tr>
<td>V</td>
<td>-</td>
<td>f,m + s,p + 1,2,3 + W, G, K, P, I, J, F, T, C, S, Y</td>
<td>Verbs</td>
</tr>
<tr>
<td>ET</td>
<td>-</td>
<td>-</td>
<td>Foreign words</td>
</tr>
<tr>
<td>PONCT</td>
<td>Strong, weak</td>
<td>-</td>
<td>Punctuation</td>
</tr>
</tbody>
</table>

Table 5. Morpho-syntactic tags.

<table>
<thead>
<tr>
<th>Phrasal category</th>
<th>Subcategorization</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;NP&gt;, &lt;/NP&gt;</td>
<td>-</td>
<td>Noun phrases</td>
</tr>
<tr>
<td>&lt;VN&gt;, &lt;/VN&gt;</td>
<td>-</td>
<td>Verbal nucleus</td>
</tr>
<tr>
<td>&lt;VP&gt;, &lt;/VP&gt;</td>
<td>-, inf, part</td>
<td>Infinitives and nonfinite clauses</td>
</tr>
<tr>
<td>&lt;PP&gt;, &lt;/PP&gt;</td>
<td>-</td>
<td>Prepositional phrases</td>
</tr>
<tr>
<td>&lt;AdP&gt;, &lt;/AdP&gt;</td>
<td>-</td>
<td>Adverbial phrases</td>
</tr>
<tr>
<td>&lt;AP&gt;, &lt;/AP&gt;</td>
<td>-</td>
<td>Adjectival phrases</td>
</tr>
<tr>
<td>&lt;SENT&gt;, &lt;/SENT&gt;</td>
<td>-</td>
<td>Sentences</td>
</tr>
<tr>
<td>&lt;S&gt;, &lt;/S&gt;</td>
<td>-, int, sub, rel</td>
<td>Finite clause</td>
</tr>
<tr>
<td>&lt;COORD, &lt;/COORD&gt;</td>
<td>-</td>
<td>Coordinated phrases</td>
</tr>
</tbody>
</table>

Table 6. Syntactic tags (simplified).
Appendix 2 Sample of the TALANA corpus

Simplified format, with morphosyntactic annotations.

Au_cours_de P+PNP
la Dfs
conférence_de_presse NCfs+NPN
qui PROR3fs
a VP3s
clos VKms
cette Dfs
rencontre NCfs
,
le Dms
premier_ministre NCms+AN
est-allemand Ams+XA
est VP3s
revenu VKms
sur P
les Dmp
incidents NCmp
de P
lundi NCms
soir NCms

SGML format (with constituency):

<SENT><PP>Au_cours_de:P
  <NP> la:Dfs conférence_de_presse:NC-fs
  <Srel><NP>:SUIJ qui:PROR-3fs</NP>
  <VN>a:VP-3s clos:VK-ms</VN>
  <NP>cette:D-fs rencontre:NC-fs</NP>
</Srel>
</NP></PP>;PONCT
  <NP>le:D-ms premier_ministre:NC-ms</NP> est-allemand:A-
  <AP></AP></NP>
  <VN>est:VP-3s revenu:VK-ms</VN>
<PP>sur:P</PP> les:D-mp incidents:NC-mp
  <PP>de:P</PP> lundi:NC-ms soir:NC-ms</PP></PP>
</SENT>
Graphical display of the same sentence:

```
SENT
  PP       NP       VN       PP
  |       |       |       |
 Au cours de le premier ministre est revenu sur
  |       |       |       |
 la conférence de presse est-allemand les incidents
  |       |       |       |
 qui a clos cette rencontre lundi soir
```

References


TALANA-UFRL
Université Paris 7
750XX Paris
France
{abeille,clement,kinyon}@linguist.jussieu.fr