Enriching a French treebank

Anne Abeillé, Nicolas Barrier

Laboratoire de linguistique formelle
Case Postale 7031
2, place Jussieu 75251 Paris Cedex 05, France
{abeille, nbarrier}@linguist.jussieu.fr

Abstract

This paper presents the current status of the French treebank developed at Paris 7 (Abeillé et al., 2003a). The corpus comprises 1 million words from the newspaper *le Monde*, fully annotated and disambiguated for parts of speech, inflectional morphology, compounds and lemmas, and 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. It has been used by computational linguists to train and evaluate taggers, parsers and lemmatizers, as well as by psycholinguists to extract lexical and syntactic preferences (Pynte et al., 2001). It is now being enriched with functional information, and used for parsing evaluation.

1. The French treebank

Similarly to the Penn TreeBank, we have annotated both parts of speech and constituents. Differently from the Penn Treebank, we have also annotated compounds, lemmas and inflectional morphology. Our annotation choices are meant to be linguistically motivated and compatible with various linguistic theories. We have chosen surface-based annotations, with no empty categories (Abeillé and Clément, 2002; Abeillé et al., 2003b; Abeillé, 2003).

With compounds amalgamated and not counting punctuation marks, the treebank comprises 870 000 tokens, using 37 000 different lemmas, making up about 32 000 independent sentences. The average number of words per sentence is 27 and the average number of phrases is 20 (some phrases are unary).

It has been automatically tagged and hand-corrected by human annotators in a first phase, and automatically chunked and hand-corrected in a second phase (Clément, 2001; Toussenel, 2001; Abeillé et al., 2003a). In the first phase, the task of the annotators was to validate the sentence boundaries, as well as the compounds (for missing compounds or possible compounds irrelevant in a given context), and to validate the morpho-syntactic tags, especially for notoriously difficult cases (for example *de* as a preposition or as a determiner). In the second phase, the annotators' task was to validate the constituant labels and boundaries, adding embedding where appropriate, as well as to signal remaining errors which could have been overlooked in the first phase. They used a specific Emacs-based annotation tool. The annotated and validated corpus is formatted in XML, using the XCES recommendations, and is available for research purposes.

We distinguish 14 lexical categories, used for simple words as well as for compounds: A (adjective), Adv (Adverb), CC (coordinating conjunction), CL (weak clitic pronoun), CS (subordinating conjunction), D (determiner) ET (foreign word), I (interjection), NC (common noun), NP (proper name), P (preposition), PRO (strong pronoun), V (verb), PONCT (punctuation mark). We distinguish 12 phrasal categories: AP (adjectival phrase), AdP (adverbial phrase), COORD (coordinated phrase), NP (noun phrase), PP (preposition phrase), VN (verbal nucleus), VPinf (infinitival clause), VPpart (participial clause), SENT (independent clause), Sint (parenthetical), Srel (relative clause), Ssub (other subordinated clause) 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 titles, it is difficult to determine the head, and we have one global NP with no internal constituents. For coordinations, we have a COORD phrase, for the conjunction and the non initial conjuncts) usually included inside a major phrase (headed by the initial conjunct). We do not have discontinuous constituents, since these can usually be recovered at the functional level : in *Combien voulez-vous de pommes* (lit. *how many do you want of apples ?*) both *combien* and *de pommes* have the same Object function.

Most of the difficult cases were with PP attachment, or scope of coordination, and human annotators had to spend the necessary time to fully understand the sentences. We got rid of spurious ambiguities (with the same interpretation) by a Attach high heuristics, for example in support verb constructions such as *écrire un livre sur les indiens* (*write a book about Indians*) where the PP complement passes the linguistic tests both as a complement of the Verb and as a complement of the preceding Noun, with no semantic difference.

2. Enrichment of the treebank

2.1. Enriching the treebank with grammatical functions

Similarly to what has been done for the German Negra or Tiger Treebanks (Brants et al., 2003), we have added some functional information to the French treebank. We chose to annotate surface grammatical functions only, and mark them as labels on the phrasal categories. For clitics, we mark the corresponding functions on the verbal nucleus. Functional information such as complement (or modifier) of Noun or complement of Adjective is already implicit in the constituent hierarchy (or in the constituent label for relative clauses). So we have concentrated on the

functional tagging of verbal dependents, for which this information was not available. We distinguish 8 grammatical functions: A-object (A-OBJ), Subject predicate (ATS), Object predicate (ATO), De-object (DE-OBJ), Direct object (OBJ), Modifier (MOD), Prepositional object (P-OBJ), Subject (SUJ).

We only annotate surface functions: the subject of passive verbs for example bears a Subject function, not an Object one. Phrases have at most one function: in case of infinitival constructions, we only note the surface function of the NP complement (with respect to the main V) and not its "deep" subject function (with respect to the infinitival V). In *Je vois Paul partir* (*I see Paul leaving*), the NP $Paul$ is annotated as the direct Object of the V $vois$, not as the Subject of the Vinf $partir$. On the other hand, two constituents can have the same function in the same sentence. It is the case with inverted clitics which are compatible with an NP subject in French. In *Paul part-il ?* (lit. *Paul does he leave ?*) both the NP $Paul$ and the following VN are tagged with a Subject function. Discontinuous dependents are another case of independent constituents tagged with the same function (such as the Object pronouns "en" and "*quelques uns*" in *On en a pris quelques uns* (lit. *We them have taken some*)). For verbal nuclei (VN), we annotate functions of the clitic pronouns included in the VN, such as Subject for "il", Direct object for "la", etc.

The grammatical functions are automatically added to the constituents (which are VN or sisters of VN) by a functional tagger developed by Jacques Steinlin and Nicolas Barrier, and then hand-corrected. It is rule-based, written in JAVA, using the XERCES API and 115 rules which are unification-based and fully ordered. The rules define underspecified patterns against which the corpus trees are matched to assign the correct function to a given constituent and allow for default assignment.

We have evaluated it against a sample of 1000 hand-corrected sentences (picked randomly from the corpus). It performs with an average precision of 89,69% (best precision for subjects: 99,47%) and an average recall of 89,27% (best recall for modifiers: 95,48%) (cf. Table 1).

Annotators are currently validating the functional tagging, using an enhanced version of our Emacs-based validation tool[1]. Human validation is significantly easier than in the previous annotations phases: only a subset of the constituents has to be considered, and it mostly involves understanding the sentence. Difficult choices imply distinguishing predicative complements from objects, and modifiers from prepositional objects. For the former, we use a list of verbs taking predicative complements, for the latter we ask the annotators to conform to linguistically available tests (modifiers are more mobile than complements, only complements can be obligatory, etc.). A distribution of the different functions among the different constituents has been computed on the same 1000 sample sentences and is presented in Table 2.

Notice that certain functions are not defined for certain constituents: no NP can be an a-object, no PP can be a subject. On the other hand, the lack of Object predicate NP in Table 2 is only due to the small size of the sample (a valid example would be *On l'a élu président* '*we have elected him president*'). More surprising cases are adjectival objects, such as "*peser lourd*" (*to weigh heavy*) or locative NPs annotated as prep-objects, such as "*aller place Beauveau*" (*to go place Beauveau*). Notice that, contrary to what is usually found in spoken French, nominal subjects are the most frequent ones (clitic subjects are annotated as VN). Notice also that adverbial phrases may be underestimated because we do not have unary adverbial phrases (we only annotate AdP with at least two elements). In case of coordination, we only annotate the embedding phrase, and not the embedded COORD. We annotate COORD phrases only when they are not embedded, that is the case with "multiple conjunctions" such as:

```
<COORD fct="SUJ">Et le Maroc</COORD>
<COORD fct="SUJ">et l'Algérie</COORD>
<VN>réussiront<VN>
```

(lit. *And Marocco and Algeria will-succeed*).

2.2. Using the enriched treebank

A small subset of the new treebank with functional information is being used in the French project EASY for parsers evaluation (Gendner et al., 2003). EASY defines a relation-based annotation scheme inspired from (Carroll et al., 2003). In order to convert our treebank into this richer format, we define a two-step conversion procedure : first our constituents are split into smaller chunks, then our functional tags (or levels of embeddings) are converted into sets of dependency relations between chunks, with a grammatical function. The first step is done automatically. The second step is performed semi-automatically with some human validation. Notice that in EASY, the functions are annotated as relations between chunks, or between words and chunks. For embedded constituents (not verbal dependents), the dependency relations can be easily read off the tree structure: a PP inside an NP for example bears a MOD_N relation with the head Noun of the NP, a PP inside an AP bears a MOD_A relation with the head Adjective in the AP etc. The only information to be added is that of headedness, and most of the time heuristics such as the first N (for a NP), the first A (for an AP) ... are sufficient. For verbal dependents, our functional tags are converted into binary relations. Long distance relations (when an NP object for example bears a relation not with the following VN but with a more distant one, as in "*Que voulez-vous dire ?*" *What do you mean to say ?*) have to be added by hand (although some automatization could be considered) as well as control relations.

Conclusion

There is a long standing debated between constituency-based annotations and dependency-based annotations for NL corpora. Dependency-based schemes are more suitable for parser evaluation purposes, whereas constituency-based ones are often more suitable for grammar extraction purposes. Although one type of annotation can in theory be converted into the other, matters are often more difficult

[1]So far, about 20% of the corpus has been validated for functional tagging.

Precision							
89,69%							
Subject	D. Object	a-object	de-object	Prep-object	S. Predicate	O. Predicate	Modifier
99,47%	92,29%	91,45%	87,93%	91,52%	95,61%	92,30%	79,04%

Recall							
89,24%							
Subject	D. Object	a-object	de-object	Prep-object	S. Predicate	O. Predicate	Modifier
92,61%	89,03%	53,76%	76,11%	56,25%	84,21%	63,15%	95,48%

Table 1: Precision and recall of the functional tagger

Subject					***Direct object***					
30,78% (1,63)					25,41% (1,34)					
COORD	NP	Ssub	VN	VPinf	AP	COORD	NP	Ssub	VN	VPinf
0,13%	80,39%	0,07%	19,31%	0,07%	0,07%	0,29%	66,59%	5,88%	6,03%	21,11%

a-object				***de-object***				***prep-object***			
3,71% (0,20)				3,74% (0,21)				1,79% (0,09)			
PP	Ssub	VN	VPinf	PP	Ssub	VN	VPinf	NP	PP	VN	VPinf
63,81%	0,50%	17,08%	18,59%	81,08%	1,49%	4,97%	12,43%	1,04%	88,54%	3,12%	7,29%

Subject predicate								***Object predicate***		
5,31% (0,28)								0,35% (0,02)		
AdP	AP	NP	PP	Ssub	VN	VPinf	VPpart	AP	PP	VPinf
1,05%	41,75%	33,68%	11,92%	4,21%	0,70%	4,91%	1,75%	57,89%	21,05%	21,05%

Modifier									
28,93% (1,53)									
AdP	AP	COORD	NP	PP	Sint	Srel	Ssub	VPinf	VPpart
2,12%	0,83%	0,83%	9,79%	59,27%	7,34%	0,51%	6,57%	6,76%	5,92%

Table 2: Functions' distribution

in practice. Our experiment in enriching our constituency-based treebank for French with functional tags shows that a hybrid treebank is a possible and useful solution to the debate.

3. References

Abeillé, A., 2003. Guide des annotations fonctionnelles. Technical report, Paris 7. `http://www.llf.cnrs.fr/Abeille`.

Abeillé, A. and L. Clément, 2002. Annotation morphosyntaxique : les mots simples - les mots composés. Technical report, Paris 7. `http://www.llf.cnrs.fr/Abeille`.

Abeillé, A., L. Clément, and F. Toussenel, 2003a. Building a treebank for French. In A. Abeillé (ed.), *Treebanks: building and using parsed corpora*. Kluwer academic publishers, pages 165–188.

Abeillé, A., F. Toussenel, and M. Chéradame, 2003b. Annotations en constituants / guide pour les correcteurs. Technical report, Paris 7. `http://www.llf.cnrs.fr/Abeille`.

Brants, T., W. Skut, and H. Uszkoreit, 2003. Syntactic annotation of a German newspaper corpus. In A. Abeillé (ed.), *Treebanks: building and using parsed corpora*. Kluwer academic publishers, pages 73–88.

Carroll, J., G. Minnen, and T. Briscoe, 2003. Parser evaluation. In A. Abeillé (ed.), *Treebanks: building and using parsed corpora*. Kluwer academic publishers, pages 299–315.

Clément, L., 2001. *Construction et exploitation d'un corpus syntaxiquement annoté pour le français*. Ph.D. thesis, Paris 7.

Gendner, V., G. Illouz, M. Jardino, P. Paroubek, L. Monceaux, I. Robba, and A. Vilnat, 2003. Proposition de protocole d'évaluation des analyseurs syntaxiques du français : PEAS. *TALN 2003*. Conférence associée : évaluation des analyseurs syntaxiques. Batz sur Mer.

Pynte, J., A. Abeillé, and F. Toussenel, 2001. Constituent length and attachment preferences in French. *AMLAP Conference*. Sarrebrucken.

Toussenel, F., 2001. Marquage de constituants sur un corpus français, résultats et exploitation linguistiques. DEA thesis, Paris 7.

```
<SENT>":PONCT
  <PP fct="MOD">Au:P<NP>début:NCms</NP></PP>,
  <VN fct="SUJ">on:CL3ms ramassait:VI3s</VN>
  <VPinf fct="OBJ">
    <PP fct="DE-OBJ">de:P<NP>quoi:PROIms</NP></PP>
    <VN>remplir:VW</VN>
    <NP fct="OBJ">quinze:DCmp sacs_poubelle:NCmp</NP>
  </VPinf>
  ":PONCT
  <Sint fct="MOD">
    <VN>indique:VP3s</VN>
    <NP fct="SUJ">Roger:NPms</NP>
  </Sint>
</SENT>
```

Appendix 1: Sample with functional annotation (annotators' format)

```
<E>
  <constituants>
    <F id="F_1">"</F>
    <groupe type="GP" id="G_1">
      <F id="F_2">Au</F>
      <F id="F_3">début</F>
    </groupe>
    <F id="F_4">,</F>
    <groupe type="NV" id="G_2">
      <F id="F_5">on</F>
      <F id="F_6">ramassait</F>
    </groupe>
    <groupe type="GP" id="G_3">
      <F id="F_7">de</F>
      <F id="F_8">quoi</F>
    </groupe>
    <groupe type="NV" id="G_4">
      <F id="F_9">remplir</F>
    </groupe>
    <groupe type="GN" id="G_5">
      <F id="F_10">quinze</F>
      <F id="F_11">sacs poubelle</F>
    </groupe>
    <F id="F_12">"</F>
    <groupe type="NV" id="G_6">
      <F id="F_13">indique</F>
    </groupe>
    <groupe type="GN" id="G_7">
      <F id="F_14">Roger</F>
    </groupe>
  </constituants>
  <relations>
     <r type="MOD_V">
        <g id="G_1"/>
        <g id="G_2"/>
     </r>
     <r type="SUJ">
        <g id="F_5"/>
        <g id="F_6"/>
     </r>
     <r type="COD-V">
        <g id="G_3"/>
        <g id="G_2"/>
     </r>
     <r type="CPL-V">
        <g id="G_3"/>
        <g id="G_4"/>
     </r>
     <r type="COD-V">
        <g id="G_5"/>
        <g id="G_4"/>
     </r>
     <r type="SUJ-I">
        <g id="G_7"/>
        <g id="G_6"/>
     </r>
     <r type="MOD_V">
        <g id="G_6"/>
        <g id="G_2"/>
     </r>
  </relations>
</E>
```

Appendix 2: Same sample in EASY format