Chapter 10

BUILDING A TREEBANK FOR FRENCH

Anne Abeillé
UFRL, Université Paris VII, FRANCE
abeille@linguist.jussieu.fr

Lionel Clément
Attol, Inria, Roquencourt, FRANCE[*]
lionel.clement@inria.fr

François Toussenel
UFRL, Université Paris VII, FRANCE
ftoussen@linguist.jussieu.fr

Abstract We present a treebank project for French. We have annotated a newspaper corpus of 1 Million words with part of speech, inflection, compounds, lemmas and constituency. We describe the tagging and parsing phases of the project, and for each, the automatic tools, the guidelines and the validation process. We then present some uses of the corpus as well as some directions for future work.

Keywords: Annotation Guidelines, Shallow Parser, French, Tagger, Tagset, Treebank

Introduction

Very few gold standard annotated corpora are currently available for French. We present a treebank project for French which started in 1997. We have annotated a newspaper corpus of 1 Million words (Abeillé & Clément 1999, Abeillé et al. 2000) with parts of speech, inflections, compounds, lemmas and constituency. We follow the EAGLES recommendations (von Rekowski 1996,

[*] Lionel Clément was at University Paris 7 while working on this project.

Ide & al. 1996, Sanfilippo & al. 1996, Kahrel & al. 1997) and develop specific annotation guidelines for French. Similarly to the Penn TreeBank (Marcus et al. 1993, Taylor et al. this volume), we distinguish a tagging and a parsing phase, and arrive at a process of automatic annotation followed by systematic manual validation and correction. Similarly to the Susanne Corpus (Sampson 1995, this volume) or the Prague treebank (Böhmova et al this volume), 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. Due to the lack of robust reusable annotation tools at the beginning of the project, we chose to develop our own, which are also presented here.

1. THE TAGGING PHASE

As reported in (Abeillé & Clément 1999), we present the general methodology, the automatic tagging phase, the human validation phase and the final state of the tagged corpus.

1.1 Methodology

Choosing the corpus. The corpus consists of extracts from the daily newspaper *Le Monde*, ranging from 1989 to 1993, and covering a variety of authors and domains (economy, literature, politics, etc.), representative of contemporary written French. It comprises roughly 1 million tokens.

Choosing the tagset. We define a complete morphosyntactic tag as follows:

1 Part of Speech (POS), for example Determiner,

2 Subcategorization, for example possessive or cardinal,

3 Inflection, for example 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 tradition (Kayne 1975), and foreign words (in quotations) which receive a special POS (ET). Punctuations 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).

We chose to annotate more than just parts of speech, for the following reasons. Some parts of speech are too inclusive (e.g. conjunctions or nouns) and further distinctions (called here "subcategories") are needed (e.g. proper and

common for nouns, subordinating or coordinating for conjunctions), if one wants to use linguistically motivated distributional classes. Some words are unambiguous for parts of speech but ambiguous for such subcategories, for example *neuf* which can either be a numeral adjective (= nine) or a predicative adjective (= new), *lui* which can either be a strong personal pronoun (= him) or a weak clitic pronoun (= to him or to her), plus can either be a negative adverb (= not any more) or a simple adverb (= more).

Inflectional morphology also has to be annotated since morphological endings are important for grouping constituents (based on agreement marks) and also because many forms in French are ambiguous with respect to mood, person, number or gender. For example, the determiner ces (these) can be either masculine or feminine, the verb form *mange* (eat) can be either indicative or subjunctive, and either first or third person, or even 2nd person imperative.

We annotate lemmas because some ambiguities remain even after tagging: *suis* is an indicative verb form 1st person singular which can correspond to the lemma *être* (be) or to the lemma *suivre* (follow), and also because it is useful for corpus queries, as well as subsequent valence annotation or word sense disambiguation Compounds also have to be annotated since they may comprise words which do not exist otherwise (e.g. *insu* in the compound preposition *à l'insu de* = unbeknownst to) or exhibit sequences of tags otherwise non-grammatical (e.g. *à la va vite* = Prep + Det + finite verb + adverb, meaning ' in a hurry'), or sequences with different grammatical properties than expected from those of the parts: *peut-être* is a compound adverb made of two verb forms, a *peau rouge* (American Indian) can be masculine (although *peau* (skin) is feminine in French) and a *cordon bleu* (master chef) can be feminine (although *cordon* (ribbon) is masculine in French). Some sequences are ambiguous between compound and literal interpretations, although in corpora the compound interpretation often prevails:

(10.1) Paul veut bien que Marie vienne (Paul indeed wants Mary to come).

(10.2) Paul pleure bien que Marie vienne (Paul is crying although Marie is coming)

(10.3) Paul en fait a raison (Paul in fact is right)

(10.4) Paul en fait trop (Paul is acting too much)

In (10.1), there is no compound: *bien* is an adverb (well) and *que* a subordinating conjunction (that); whereas in (10.2) the same sequence *bien que* is a compound subordinating conjunction (although). In (10.3), the sequence *en fait* is a compound adverb (in fact), whereas in (10.4) the same sequence must be decomposed into *en* as a clitic and *fait* (does) as a finite verb.

Compounds are annotated with the same tagset as non-compounds, but tags are added for each of their parts. Since the borderline between compounds and free sequences is subject to much linguistic debate, we chose to also annotate the parts of the compounds. Tagging the parts as well is useful for specific studies on compounds, but also if a user wants to view our corpus ignoring some or all the compounds.

Our complete tagset comprises 218 tags, which are valid combinations of the labels presented in Table 10.1.

Table 10.1. Tagset of the tagged corpus

POS	**Subcategory**	**Morphology**	**Description**
A	cardinal, ordinal, poss, qualif., indef, inter	f,m + s,p + 1,2,3	Adjectives
Adv	-, inter, exclam, negative		Adverbs
CL	subj, refl, obj, -	f,m + s,p + 1,2,3	Clitic pronouns
C	subord, coord	-	Conjunctions
D	card, dem, def, indef, exclam, negative, poss, inter, partitive	f,m + s,p + 1,2,3	Determiners
ET	-	-	Foreign words
I	-		interjections
N	common, proper	f,m + s,p	Nouns
P	-	-	Prepositions
PRO	inter, pers, card, neg, poss, rel, indef	f,m + s,p + 1,2,3	Other Pronouns
PONCT	strong, weak	-	Punctuation
PREF	-	-	Prefixes
V	-	f,m + s,p + 1,2,3 + W, G, K, P, I, J, F, T, C, S, Y	Verbs

The tagging pipeline. The overall organization of the tagging phase (cf. Figure 10.1) is more complex than just automatic tagging followed by human validation, because of the rich tagset we are using. For practical reasons, segmentation (for compounds) was done before tagging and lemmatization after tagging. For each phase, we try to minimize the number of tags involved, so in practice we define three different tagsets: a reduced one for the tagger (in order to minimize its errors), an enriched one for the annotators (so that all possible ambiguities are resolved but annotators don't have to make distinctions that are easy to perform automatically), and the full tagset for final formatting of the treebank. Mapping tools between these tagsets have thus been developed.

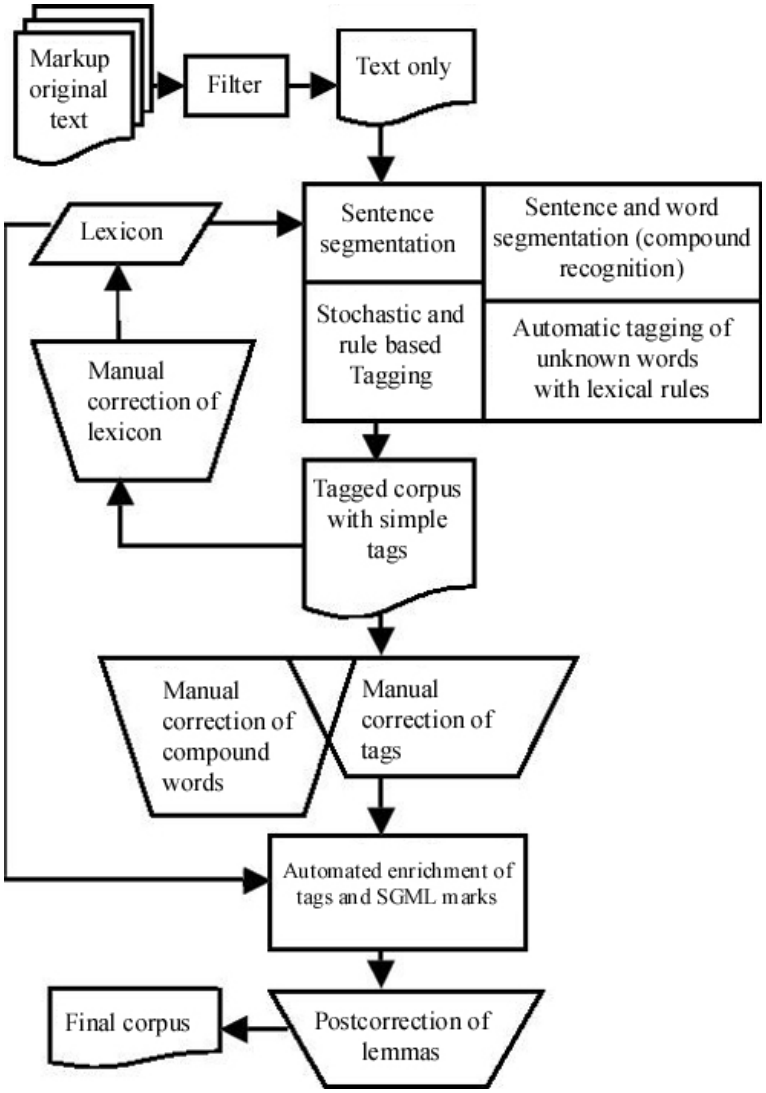

Figure 10.1. The tagging pipeline

1.2 Automatic tagging of the corpus

Given the lack of reusable annotation tools at the beginning of the project, we developed a morphosyntactic tagger for French (Reyes 1997, Clément 2001). Our tagger is based on Brill (1993) in that it has two phases (initial or dummy tagging, and context sensitive tag rewriting). The main difference with respect to a true Brill tagger is that we have added an external lexicon and rely mainly on manually written contextual rules. The tagger uses a reduced tagset for POS and morphology (103 tags), mainly derived from existing lexicons, with only a few simplifications for distinctions difficult to handle automatically (for example between interrogative and relative pronouns which are ambiguous forms in French).

The tagger's lexicon. The lexicon of the tagger comprises over 360,000 forms including 36,000 compounds. It comes from the lexicon we developed for our French Tree Adjoining grammar (FTAG, Candito 1999, Abeillé 2002), plus some extracts from the MULTEXT lexicon, from ABU lexicon (for proper names) and from INTEX for compounds (Silberztein 1993). It has been extended with most forms from the corpus (excluding numbers and specific proper names).

Segmentation. Our tagger comprises a sentence splitter and a tokenizer. The sentence splitter uses lexical data to properly distinguish between capital letters for proper names or beginning of sentence, between dots for acronyms or end of sentence, between hyphenation and linking dashes, etc. (cf. Silberztein 1993). As in English, word segmentation is always a problem since many compounds show up as separate words in French (*pomme de terre* = potato, *bien que* = although, etc.) The tokenizer thus reads the lexicon for compound recognition and groups the best compound candidates (choosing the longest one in case of several candidates; for example the compound adverb (or noun) *face à face* and not the compound preposition *face à* in the sequence *face à face*). Grouping compounds before tagging helps the tagger in most cases and leads to a few errors, which are corrected in the validation phase[1].

Unknown words. As with Brill's tagger, our tagger uses lexical rules for unknown words. We currently have 198 such rules corresponding to common suffixes for verbs, nouns, and adjectives. They are somewhat similar to a morphological analyzer. We also have regular expressions for numbers and proper nouns (acronyms). Since our lexicon has been extended as part of the project, most unknown words are foreign words and typographical errors.

Initial tagging. The initial (dummy) tagging is important since more than 40% of the words in our corpus receive more than one tag with the tagger's lexicon (and more than 20% more than two tags). In order to assign the best possible tag for each word, we rely on the trigram method using genotypes as data (cf. Tzoukerman 1995) and compute the probability of each tag for each word with a large unannotated corpus (of newspaper texts)[2].

Contextual retagging. The initial tag assigned to each word (by lexical lookup) can be changed depending on the context. The form *été* is assigned "Verb" (past participle) as its most probable tag, but must be retagged as Noun after a Determiner for example. Contrary to Brill's automatic rule induction approach (which gave poor results, cf. Reyes 1997, Abeillé & Clément 1999), we preferred to develop most of the contextual rules by hand, based on linguistic knowledge and corpus lookup. In order to accomodate the linguist's need for expressivity, we have added compositionality to the rule formalism, as well as three operators (for negation, for testing whether a tagset contains a specific tag, and to force a transformation even if a tag is not in the tagset of the word in the lexicon). We also added the possibility of having unifiable variables in the tag names (for morphological agreement). The tagger uses 322 contextual (retagging) rules[3].

The contextual rules cannot deal with compound/literal ambiguity (*carte bleue*: blue card or credit card) since they cannot modify the initial segmentation. The tagger does not handle lemmas, which are handled by a specific postprocessor. It processed our corpus with an error rate of about 8%.

1.3 Validating the tagged corpus

Word segmentation (compound recognition) had to be validated by systematic human scrutiny of the tagged corpus, as well as for each form (simple or compound) POS and inflection. Some subcategorization information, most lemmas and all parts of compounds depend only on lexical information (independently of the context) and were added automatically by lexical lookup (once compound segmentation, POS and morphology had been manually validated).

For the two main manual validation tasks (compound validation and tagging validation), very precise guidelines were written (Abeillé et al. 2000) and updated during the project. The reference tagged corpus was checked and corrected by two annotators (one after another) reading (some part of) the text in a longitudinal way, then some tools were applied to check the most difficult cases (for some frequent grammatical words such as *de* or *que*). Weekly meetings were also necessary to ensure consistency between annotators (more than fifteen persons were involved altogether). There were no remaining ambiguities at this level.

Validation of compounds. The automatic annotation for compounds was done by INTEX (Silberztein 1993) and by our tagger. We asked the annotators to validate the compound interpretation in context, to add compounds missing from our lexicons (especially for proper names) and to add discontinuous instances of compounds that could not be automatically found. We gave them guidelines about what to consider a compound based on linguistic tests (cf. Gross 1996), using morphological tests (parts not existing otherwise: *fi* in *faire fi de* (scorn)), syntactic tests (no internal modification or determination : *carte bleue* (credit card) / **carte très bleue* (very blue card) and semantic criteria such as opacity : en revanche (on the contrary; literally: in a revenge)). The well-known preference for compound interpretation was confirmed, except for function words, for which lots of candidate compounds turned out not to be compounds. For example *de plus* can be the compound adverb (all the more) but most of the time it was the preposition (*de*) followed by the adverb (*plus*). To our surprise, very few discontinuous compounds (*afin<justement>de* 'in order precisely to') were found in the corpus.

Annotating the parts of the compounds was done automatically with our compound lexicon.

Validation of tags. Our complete tagset comprises 218 tags (see Figure 10.1). Since most subcategories can be assigned unambiguously to a word (once its POS is known), we chose to simplify the tagset for the annotators. Possessive determiners, for example, can be ambiguous with other POS but not with other determiner subcategories, and the same for possessive pronouns; so the subcategory Possessive can be eliminated from the annotators' tagset. The tagset for the annotators was thus reduced to 122 tags, and they were presented with subcategories only in case of possible ambiguity (*neuf* as cardinal or qualifying adjective ('nine' or 'new'), *lequel* ('which') as interrogative or relative pronoun etc.). Annotators had to validate the output of the tagger and to add subcategories when needed. Most of the subcategories were added automatically afterwards with lexical lookup.

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 Multext 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 / reflexive subcategories, since assigning the right 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 has to find their antecedent to properly annotate their morphosyntax.

Most of the 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).

Validation of lemmas. The lemmas were not shown to the annotators. They were added automatically (using our lexicon) after the tag correction phase. At this stage, very few lemma ambiguities remained (*fils* NCmp from *fil* 'thread' or *fils* 'son' ...). They were resolved by hand. Some well known ambiguities such as *savons* (= *savon* 'soap' or *savoir* 'know'), *portes* (= *porte* 'door' or *porter* 'carry'), which are problematic for most lemmatizers, do not arise once the corpus has been tagged with parts of speech. The cost of validating 1 Million words was about 50 man-months (including tagger development). The average correction speed was 500 words per hour. This is much slower than that of the Penn Treebank (2000 words per hour) because of the compounds, and because of our richer tagset (the annotators were presented only 36 tags in the Penn Treebank).

2. THE PARSING PHASE

After the tagging phase was completed, we started the parsing phase of the project. We first present our annotation choices, then our tools for automatic syntactic annotation, and the validation phase. Human validation of constituency is now complete; current and future work includes functional annotation and XML formatting.

2.1 Syntactic annotation scheme

Contrary to tagging, precise language-specific guidelines are usually missing for syntactic annotation. In order to provide annotations that are reusable by researchers with various backgrounds, we chose to annotate both constituency and functional relations, starting with constituency.

Choosing the tagset. 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 Possible subcategory (e.g. Rel for relative clauses)

3 Surface function (eg. Subj, Object for NPs)

4 Opening or closing boundaries (<>, </>)

5 Diathesis (e.g. passive) for verbal nuclei

For the moment we have only annotated phrasal names (category and subcategory) and phrasal (i.e. constituent) boundaries. The current syntactic tagset is shown in Table 10.2:

Table 10.2. Tagset of the parsed corpus

Phrasal category	Subcategorization	Description
<NP>, </NP>	-	Noun phrases
<VN>, </VN>	-	Verbal nucleus
<VP>, </VP>	-, inf, part	Infinitives and nonfinite clauses
<PP>, </PP>	-	Prepositional phrases
<AdP>, </AdP>	-	Adverbial phrases
<AP>, </AP>	-	Adjectival phrases
<SENT>, </SENT>	-	Sentences
<S>, </S>	-, int, sub, rel	Finite clauses
<COORD>, </COORD>	-	Coordinated phrases

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 parsimonious use of unary phrases: we have unary NPs for proper names and pronouns, but not for bare common nouns, we have unary APs for predicative adjectives, but not for modifying ones, we have unary VNs for verbs but no unary AdP for Adverbs (see Abeillé et al. 2000). For rigid sequences of categories, such as dates or addresses, it is difficult to determine the head, and we have one global NP with no internal constituents.

We annotate certain phrases with a subcategory, which is important for functional annotation, for example relative or subordinate for embedded clauses.

We do not have discontinuous constituents. For discontinuous objects such as '*en ... trois*' in (10.5) or '*combien ... de pommes*' in (10.6), we will mark that they share the same Object function at the functional level, since a constituent made of a clitic (*en*) and a pronoun (*trois*) would not make sense from the syntactic point of view.

(10.5) `<NP> Paul:NP </NP> en:CL mange:V <NP> trois:PRO </NP>`

(Paul eats three of them)

(10.6) `Combien veux-tu <NP> de pommes </NP> ?`

(how many apples do you want ?)

In order to be as theory neutral as possible, we neither use empty categories, nor functional phrases (no DP or CP). We allow for headless phrases (elliptical NP lacking a head Noun as in (10.7) or sentential clauses lacking a verbal nucleus as in (10.8)).

(10.7) `Ce sont <NP> les:D meilleurs:A </NP>`

(they are the best)

(10.8) `plus vieille <Ssub> que:CS <NP> toi:PRO </NP> </Ssub>`

(older than you)

Unexpressed subjects (in infinitives or participials) will be marked at the functional level.

We made two specific choices, regarding verbal phrases, and regarding coordinated phrases, in accordance with some particularities of the French language.

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 (like in Italian, cf Montemagni et al this volume):

(10.9) Les actions **qu'a mises** IBM **sur le marché** (the shares that IBM put on the market)

(10.10) Les actionnaires **décideront** certainement **une augmentation de capital** (the stock holders will certainly decide on an increase in capital)

In (10.9) the NP subject (IBM) is postverbal and precedes the locative complement (*sur le marché*). In (10.10), the adverb *certainement* is also postverbal and precedes the NP object (*une augmentation de capital*).

For coordination, we only mark a Coordinating phrase after a coordinating conjunction. We do not necessarily embed conjuncts inside a phrase since there

are cases where there is none (10.11), and there are cases where the category of the phrase would be unspecified (10.12):

(10.11) `Paul va <NP>le lundi</NP> <PP>à <NP>la piscine</NP> </PP> <COORD> et <NP>le mardi</NP> <PP>au <NP>cinéma</NP></PP> </COORD>.`
(*Paul goes on Monday to the swimming pool and on Tuesday to the movies*)

(10.12) `Paul est <NP>médecin</NP> <COORD> et <AP>fier <VPinf>de l'être</VPinf></AP></COORD>`
(*Paul is a doctor and proud of it*)

We consider the first conjunct as the head and annotate each following conjunct with a specific category COORD[4].

The parsing pipeline. The following gives an overall view of the complete parsing phase[5]:

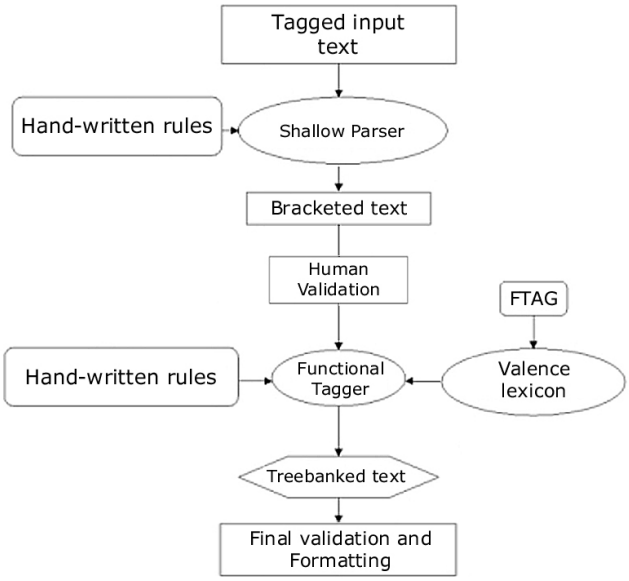

Figure 10.2. The parsing pipeline

2.2 Automatic parsing of the tagged corpus

We choose to use different tools for each task. We need a robust parser for marking major phrase boundaries and a functional tagger for marking syntactic functions (on major phrases) and valence (for each main verb).

For marking constituency, we do not use a classical parser, but have instead adapted specific tools more robust and more suited to our goal. We use a rule-based shallow parser (Kinyon 2001) which marks major phrase boundaries (NP, PP,), with limited embedding (and no recursion) (cf Abney 1991, Giguet 1998). It is designed to minimize errors, so it does not try to attach PPs or relative clauses. These attachments have to be added by the human annotators.

The shallow parser was first developed by Kinyon (2001) and adapted to the corpus by Toussenel (2001). We chose a rule-based parser which does not need a pre- existing treebank (contrary to probabilistic ones.

The shallow parser takes as input the tagged text, slightly simplified (discarding the lemma and the morphological information, but retaining some POS subcategories). It adds phrase boundaries in a left-to-right fashion. It was developed in Java for portability and currently comprises approximately 40 rules. Each rule has access to a limited context : the previous, current and next tag plus the label of the constituent(s) currently being processed. The main underlying idea is to rely on function words as triggers of constituent boundaries (e.g. when encountering a determiner, start a noun phrase (NP), or when encountering a clitic, start a Verbal nucleus)[6].

The shallow parser uses the same tagset as that in Table 10.2, except for Sint, which has to be added by the annotators, and INC (for unknown constituents), which is replaced in a postprocessing phase.

Following linguistic tradition, we consider as function words all words associated with a POS that labels a closed class : i.e., determiners, prepositions, clitics, pronouns (relative, demonstrative), conjunctions (subordination and coordination), auxiliaries, punctuation marks and adverbs that belong to a closed class (e.g. negation adverbs *“ne”, “pas”*).

The general idea is that when one of these function words is encountered, an opening boundary for a new constituent is inserted in the text. Closing boundaries are added either naturally when a new constituent begins (e.g. NPs end when a new constituent starts), or triggered by a new function word (e.g. relatives and sentential complements end when a punctuation mark or a conjunction is encountered). Of course, some rules may refer to non function words (e.g. when encountering a proper noun, start an NP).

An example of a rule is the following:

```
When the current word is a preposition, look at the next word.
If this is a verb or a clitic or an adverb, then
     start <VPinf>.
Else start <PP>.
```

(First case) *le temps* <***VPinf***> ***de*** <*VN*> *parler* </*VN*> *soigneusement*
('the time for talking carefully')

(Second case) <***PP***> ***de*** <*NP*> *Moscou* <*NP*> <*PP*>
('from Moscow')

The shallow parser yields an output in linear time, since the input text is scanned just once, and constituent boundaries are added incrementally in a monotonic manner. Contrary to (Abney 1991), it does not establish a link between the constituent boundaries and the prosodic pattern in sentences. Contrary to (Ait- Mokhtar & Chanod 1997), it works in a strictly left-to-right fashion, and performs limited embedding (NPs embedded inside PPs , VN embedded inside a relative clause, itself embedded inside an NP) and limited attachment (e.g. coordination).

An evaluation has been done automatically by comparing the output of the parser with the corrected version of a set of 500 sentences (picked randomly from the corpus). For opening brackets, the recall is 92.9% and the precision is 96%, with 94% correct labels. For closing brackets, the recall is 58% and the precision is 60%, with 56.5% correct labels (cf. Toussenel 2001). Errors difficult to correct with access to a limited context involve mainly "missing" brackets (e.g. "*comptez-vous *ne pas le traiter*" appears as a single constituent, while there should be 2), while "spurious" brackets can often be eliminated by adding more rules (e.g. for multiple prepositions: "*de chez*"). For closing brackets, most of the errors are due to unattached constituents.

2.3 Validating the parsed corpus

For human validation, we defined precise guidelines, drawing on grammar books (cf. Grevisse & Goosse 1992) as well as numerous syntactic studies, especially in surface-based frameworks (cf Abeillé & Godard 2000). The annotators' task consists of the following steps:

1 checking (and enriching) the names of the syntactic tags,

2 checking (and possibly moving) the position of the syntactic tags.

The first check is usually done manually only on opening boundaries and the second check usually involves moving closing boundaries.

We are checking in a longitudinal way using an Emacs-based tool especially designed by Michel Simard and Lionel Clément called CAT. The tool presents the annotator with one phrase per line with automatic indentation according to the level of embedding. It enables the annotator to move opening or closing phrase boundaries, to view the sentences as a graphical tree, to change automatically opening and closing boundary names, and to view or hide full morphosyntactic tag of each word.

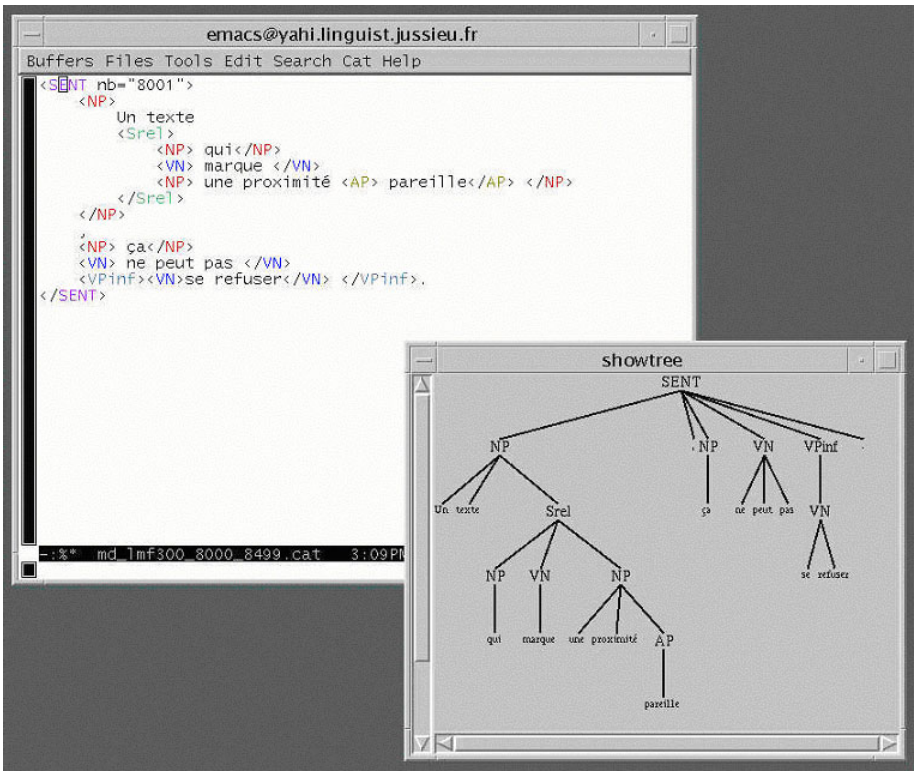

Figure 10.3. The CAT Validation tool for phrases

Most of the difficult cases involved PP attachment, Relative clause attachment or scope of coordination, for which a deep understanding of the sentences is necessary. We were agnostic about ambiguities remaining at the beginning of the project. But most of the ambiguities were resolved at the semantic level. The remaining ambiguities such as (10.13) are thus only spurious ones and we chose to get rid of them by the "attach high" heuristics (namely the second parse here).

(10.13) Paul écrit un livre sur les Indiens (Paul writes a book about Indians)

```
<NP>un livre <PP>sur <NP>les Indiens</NP></PP></NP>
```

or

```
<NP>un livre </NP> <PP>sur <NP>les Indiens</NP></PP>
```

Some mislabeled constituents were those introduced by other function words than those defined in the parser's rules, eg; NPs starting with quantificational adverbs or prepositions:

(10.14) `<NP>environ:ADV trois cents:DET personnes :N</NP>`
(approximately 300 people)

(10.15) `Ils achetent <NP>jusqu'à:P 10 000:DET actions:N</NP>`
(they buy up to 10,000 shares)

Some unexpected parsing errors came from remaining tagging errors that had escaped previous human validation. We checked using sample texts that different annotators were converging on the same parse, and usually one postchecking was sufficient for ensuring the quality of constituency annotation.

2.4 Syntactic formatting and search tools

The tagged corpus has been formatted in XML following the TEI and XCES guidelines (cf. Ide et al. 1996; see Appendix 2). The parsed corpus will be formatted in XML too (cf Ide and Romary this volume).

A search tool has been developed by L. Clément (Clément 2001) with a graphical interface, for linguists looking at specific words, expressions or constructions in French. This tool will be developed for syntactic annotation, too.

3. CURRENT STATE AND FUTURE WORK

3.1 Current state of the treebank

The automatically tokenized, tagged and parsed corpus has been manually validated and enriched (with longitudinal exhaustive checking by at least three different human annotators). The 1 Million tokens when annotated amount to 870,000 words, excluding punctuation signs, and clustering compounds into one word, for a total of 32,000 sentences with 37,000 different lemmas. The average number of words per sentence is 27 and the average number of internal constituents is 20. There are no remaining ambiguities, nor unknown words (the original typos have been corrected).

It is available in two versions :

- a light version with a reduced tagset in a compact format,
- a full version with a richer set of tags, lemmas, constituents, all in SGML format (see appendix).

3.2 Using the treebank

The treebank is freely available for research purposes and is currently used by more than 10 licensees[7].

The tagged version has been used for evaluating a lemmatizer (Namer 2000), and for training stochastic taggers (cf. Langlais & Simart 2002). L. Clément used it to train a tagger based on Treetagger (Schmid 1994) as part of the RLT project. The parsed version has been used in collaboration with psycholinguistics at LPL (Aix) to check syntactic preferences (Abeillé et al 2001) and to train a connectionnist parser. Our guidelines have been reused for other projects (CLIF, Multitag) and will be used for the EASY project of parser evaluation for French (which is inspired by Carroll et al. this volume).

3.3 Future work

The treebank will be enriched with grammatical functions, associated with the major constituents. J. Steinlin has developed a rule-based functional tagger, which has to be fine-tuned to the corpus, and enriched with a valence lexicon, such as that developed within the FTAG project (cf. Barrier 1999). With such information, an automatic conversion of the constituent structures into dependency structures will also be possible.

4. CONCLUSION

We have presented a treebank project for French. We have developed a 1 M. word reference corpus for French (from newspaper texts) and annotated it with morphosyntax, lemmas, compounds and phrase boundaries. The automatic annotations have been validated by human annotators in several steps. The reference tagged corpus is a resource to be distributed in two versions: light (with a reduced tagset) or complete (with lemmas, constituent boundaries and full SGML markup). As part of the project, several resources have been developed:

- a tagger inspired by (Brill 1993) but with mostly hand-written rules, a sizeable external lexicon and a tokenizer (handling compounds),
- a shallow parser marking major phrase boundaries with limited embedding,

- guidelines for annotating constructions usually overlooked in grammar books (dates, measures, direct speech, unbalanced coordination)

The next step will be to annotate our corpus for functional relations, and to enrich our search tools. A longer term objective is the marking of some anaphoric relations (for pronouns) and of valence for verbs.

Acknowledgments

This project was mainly sponsored by the Institut Universitaire de France (IUF), but also by the CNRS (as part of the CLIF project), by LORIA (as part of the CALIN project) and by Aupelf-Uref (as part of the Francil Project). Many thanks are due to many people, including Claire Blanche-Benveniste, Nicolas Barrier, Martine Chéradame, Alexandra Kinyon, Patrick Paroubek, Rodrigo Reyes, Laurent Romary, Jacques Steinlin, Michel Simart, Jean Véronis, Eric Wehrli, and the annotators themselves.

Notes

1. We also did some lexicon tuning, taking out the compound adverb *sur ce* (colloquial 'then') for example since the corresponding sequence (Prep Det) was never a compound in the whole corpus.

2. The genotype of a word is the set of POS tags that can be assigned to it, i.e. [N:ms A:ms] is the genotype of the word /droit/ (right). [N:ms A:ms] and [A:ms N:ms] are the same genotype. Tzoukermann shows that words that share the same genotype also share (roughly) the same distribution.

3. INaLF's tagger, also based on Brill's tagger, uses about 335 rules, cf. (Lecomte & Paroubek 1996).

4. In final formatting, COORD can be replaced with the category of the first conjunct and Coord as a subcategory.

5. The last step with the functional tagger is currently being developed by J. Steinlin and N. Barrier.

6. In psycholinguistics, the role of function words in human sentence processing has been emphasized as early as (Kimball 1973), with the "new node principle" stating that "The construction of a new node is signaled by the occurrence of a grammatical function word".

7. We currently distribute one third (300,000 words= 10,000 sentences) of the corpus.

References

A. Abeillé (2002). *Une grammaire électronique du français*. CNRS Editions, Paris.

A. Abeillé, L. Clément (1999). A tagged reference corpus for French, *LINC'99 Proceedings*, EACL workshop, Bergen.

Abeillé A., D. Godard (2000). French word order and lexical weight, in R. Borsley (ed) *The Nature and Function of Syntactic Categories*, Syntax and Semantics, Academic Press, 325-358.

A. Abeillé., L. Clément, M. Chéradame, F. Toussenel (2000). Guides pour les Annotateurs - Mots simples - Mots composés - Constituants - Technical Reports, University Paris 7.

A. Abeillé, L. Clément, A. Kinyon, F Toussenel, (2001). The Paris 7 annotated corpus for French : some experimental results, in Wilson (ed) *Corpus Linguistics*, Lancaster.

A. Abeillé, J. Pynte, F. Toussenel, (2001). Constituent length and attachment preferences in French, *AMLAP*, Sarrebruck.

S. Abney (1991). Parsing by chunks. In Berwick et al (eds), *Principle-based Parsing*. Kluwer.

S. Aït-Mokhtar, J.P. Chanod (1997). Incremental Finite-State Parsing. *Proceedings of ANLP'97*, Washington, p. 72-79.

S. Barrier (1999). Repérage et classification de valences verbales, expériences avec FTAG, DEA thesis, University Paris 7.

A. Böhmova et al. (2003). The Prague Dependency treebank, this volume.

T. Brants, S. Skut, H. Uszkoreit (2003). Syntactic annotation of a German newspaper corpus, (this volume).

E. Brill (1993). A corpus-based approach to Language Learning, Ph.D. Dissertation, Department CIS, University of Pennsylvania.

J. Carroll, G. Minnen, T.Briscoe (2003). Parser evaluation using a grammatical relation annotation scheme (this volume)

M-H. Candito (1999). Représentation hiérarchique de grammaires lexicalisées: application au français et à l'italien, PhD Thesis, University Paris 7.

L. Clément (2001) Construction et exploitation d'un corpus syntaxiquement annoté pour le français, PhD Thesis, Paris 7.

S. Colonna (2001). Facteurs influençant la levée des ambigutés syntaxiques, PhD thesis, University of Provence.

E. Giguet 1998. Méthodes pour l'analyse automatique de structures formelles sur documents multilingues. PhD Thesis. Université de Caen.

M. Grévisse, A. Goosse (1992) *Le bon Usage*, Liège: Duculot.

G. Gross (1996). *Les expressions figées*, Ophrys, Gap.

N. Ide, J. Veronis, G. Priest-Dorman. (1996). Corpus Encoding Standard. EAGLES/MULTEXT.

N. Ide, L. Romary (2002) Encoding syntactic information, in this volume.

P. Kahrel, R. Barnett, G. Leech. (1997). Towards cross-linguistic standards or guidelines for the annotation of corpora. In *Corpus annotation*. Garside & al (eds). p. 231-243. Longman.

R. Kayne (1975). *French Syntax: the transformational cycle*. Current Studies in linguistics. MIT press.

J. Kimball (1973). Seven Principles of surface structure parsing in natural language. *Cognition 2*(1).

A. Kinyon (2001). A language independent shallow parser compiler. *Proceedings ACL-EACL Conference*, Toulouse.

P. Langlais, M. Simart (2002) Merging example-based and statistical machine translation: an experiment, *Proceedings of the 5th AMTA Conference*, Tiburon.

J. Lecomte, P. Paroubek (1996). Le catégoriseur d'E. Brill: mise en œuvre d'une version entraînée pour le français, Technical report, INaLF,Nancy.

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

F. Namer (2000). FLEMM : un analyseur flexionnel du français à base de règles, *TAL*, p. 41:2, p. 523-548.

U. von Rekowski (1996). ELM-FR : Specifications for French morphosyntax, lexicon specification and classification guidelines, EAGLES document.

R. Reyes (1997). Un Etiqueteur du français inspiré du taggeur de Brill, Rapport de stage, UFRL, Paris 7.

G. Sampson (1995). *English for the computer*. Oxford University Press.

G. Sampson (2003). Thoughts of two decades of drawing trees, in this volume.

A. Sanfilippo et al. 1996. EAGLES Subcategorization Standards. `http://www.icl.pi.cnr.it/EAGLES96/syntax/syntax.html`

H. Schmid (1994). Probabilistic part-of-speech tagging using decision treees, *Proceedings International Conference on new methods in Language processing*, Manchester.

M. Silberztein (1993). *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*, Paris : Masson.

B. Srinivas, A. Joshi (1999). Supertagging: An approach to almost parsing, *Computational Linguistics*, 25-2: 237-266.

A. Taylor, M. Marcus, B. Santorini (2003). The Penn treebank: an overview, in this volume.

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

E. Tzoukermann, D. Radev, W. Gale (1995). Tagging French without lexical probabilities - combining linguistic knowledge and statistical learning, *Proceedings EACL SIGDAT Workshop*, Dublin.

Appendix

1. Sample extract

Sample text
Au cours de la conférence de presse qui a clos cette rencontre , le premier ministre est-allemand est revenu sur les incidents de lundi soir.
(During the press conference which ended this meeting, the East German Prime minister talked about Monday evening)

Tagged corpus (Light format: limited tagset, no lemmas) = Annotators' format

```
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
```

2. Tagged corpus (SGML final format, simplified for compounds)

cc: tags of parts for compounds

```
<w lemma=''au cours de'' cat=P cc=PCD >
     <w lemma=''à'' cat=''P''>Au</w>
     <w lemma=''le'' cat=''D''></w>
     <w lemma=''cours'' cat=''N''>cours</w>
     <w lemma=''de'' cat=''P''>de</w>
     </w>
<w lemma=''le'' cat=D subcat=def mph=fs >la</w>
<w lemma=''conference de presse'' cat=N subcat=comm mph=fs cc=NDN >
     <w lemma=''conférence'' cat=''N''>conférence</w>
     <w lemma=''de'' cat=''P''>de</w>
     <w lemma=''presse'' cat=''N''>presse</w>
```

```
     </w>
<w lemma=''qui'' cat=PRO subcat=R mph=fs >qui</w>
<w lemma=''avoir'' cat=V mph=P3s >a</w>
<w lemma=''clore'' cat=V mph=Kms >clos</w>
<w lemma=''ce'' cat=D subcat=dem mph=fs >cette</w>
<w lemma=''rencontre'' cat=N subcat=comm mph=fs >rencontre</w>
<w lemma='','' cat=PONCT >,</w>
<w lemma=''le'' cat=D subcat=def mph=ms >le</w>
<w lemma=''premier_ministre'' cat=N subcat=comm mph=ms cc=AN >
     <w lemma=''premier'' cat=''A''>premier</w>
     <w lemma=''ministre'' cat=''N''>ministre</w>
     </w>
<w lemma=''est-allemand'' cat=A mph=ms cc=XA >
     <w lemma=''est'' cat=''A''>est</w>
     <w lemma=''allemand'' cat=''A''>allemand</w>
</w>
<w lemma=''être'' cat=V mph=P3s >est</w>
<w lemma=''revenir'' cat=V mph=Kms > revenu</w>
<w lemma=''sur'' cat=P > sur</w>
<w lemma=''le'' cat=D subcat=def mph=mp > les</w>
<w lemma=''incident'' cat=N subcat=comm mph=mp> incidents</w>
<w lemma=''de'' cat=P > de</w>
<w lemma=''lundi'' cat=N subcat=comm mph=ms > lundi</w>
<w lemma=''soir'' cat=N subcat=comm mph=ms > soir</w>
```

3. Parsed corpus (simplified, without function annotation)

```
<SENT><PP>Au_cours_de:P
       <NP> la:Dfs conférence_de_presse:NC-fs
            <Srel> <NP>:SUJ 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<AP>est-allemand:A-ms</AP></NP>
<VN>est:VP-3s revenu:VK-ms</VN>
<PP> sur:P <NP>les:D-mp incidents:NC-mp
        <PP> de:P <NP>lundi:NC-ms soir:NC-ms</NP></PP>
        </NP> </PP>
</SENT>
```

With graphical display :