Tagging French text: tagset, dedicated lexicons and guesser

Jean-Pierre Chanod and Pasi Tapanainen
Rank Xerox Research Centre, Grenoble Laboratory
6, chemin de Maupertuis, 38240 Meylan, France

Abstract

We have earlier described two taggers for French, a statistical one and a constraint-based one, in (Chanod and Tapanainen, 1995). The two taggers have the same tokeniser and morphological analyser. In this paper, we describe aspects of this work dealing with the definition of the tagset, the building of the lexicon derived from an existing two-level morphological analyser and the definition of a lexical transducer for guessing unknown words.

1 Background

We have earlier described two taggers for French: the statistical one has an accuracy of 95–97% and the constraint-based one 97–99% (see (Chanod and Tapanainen, 1995)). The disambiguation has been described before, and we discuss here the other stages of the process, namely definition of the tagset, transforming a current lexicon to a new one and guessing the words that do not appear in the lexicon.

Our lexicon is based on a finite-state transducer lexicon (Karttunen et al., 1992). The French description was originally built by Annie Zaenen and Carol Neidle, and later refined by Jean-Pierre Chanod (1994). Our lexicon is based on this transducer. The lexicon was not originally designed for a (statistical) tagger and it contains much information that is not appropriate for the tagger.

2 Tagset

2.1 The size of the tagset

Our basic French morphological analyzer was not originally designed for a (statistical) tagger, and the number of different tag combinations is quite high in it. The size of the tagset is only 88; but because a word is typically associated with a sequence of tags, the number of different combinations is higher, 353 possible sequences for single French words. If we also consider words joined with an article or a clitic pronoun, the number of different combinations is much higher, namely 6525.

If we are using a rule-based or a constraint-based tagger, the big tagset does not itself cause trouble because one can refer to a combination of tags as easily as to a single tag. On the contrary, a bigger tagset may introduce information that helps to disambiguate more accurately.

For a statistical tagger a big tagset may be a major problem. If we have 6525 tags we would need at least 40 million word corpus before we theoretically get all the two-tag combinations, and depending on the training corpus we may have different distribution for, say, present and past tenses whereas syntactically they should be like a similar way.

On the other hand, distinctions between present and past tenses, or indicative and subjunctive moods are extremely difficult to resolve by a part-of-speech disambiguator.

Therefore we used two principles for forming the tagset: (1) the tagset should not be big and (2) the tagset should not introduce distinctions that are not resolvable in this level of analysis.

2.2 Reduced tagset

There are a couple of guidelines that we used to reduce the tagset. The first one is that the tags should refers to syntactic categories, i.e. if two words have the same tag, the distributional behaviour should be similar.

2.2.1 Verb tense and mood

A guideline was that distinctions that cannot be resolved at this level of analysis should be avoided. We do not have information about the tense of the verbs. Some of this information can be recovered
later by making another lexicon lookup after the analysis. Thus if the verb tense is not ambiguous, we have not lost any information, and even if it is, a part-of-speech tagger could not resolve the ambiguity very reliably anyway. For instance, \textit{dort} (present; \textit{sleeps}) and \textit{dormira} (future; \textit{will sleep}) have the same tag \textit{VERB-SG-P3}, because they are both singular, third-person forms and they both can serve as the main verb of a clause. If needed, after the disambiguation, we can do another lexicon lookup for words that have the tag \textit{VERB-SG-P3} and assign a tense to them. Thus the tagset and the lexicon together may make finer distinctions.

On the other hand, the same form \textit{dit} can be third person singular present indicative or third person singular past (passé simple) of the verb \textit{dire} (to say). We do not introduce the distinction between those two forms, both tagged as \textit{VERB-SG-P3}, because determining which of the two tenses is to be selected in a given context goes beyond the scope of the tagger. However, we keep the distinction between \textit{dit} as a finite verb (present or past) and as a past participle, because this distinction is properly handled with a limited contextual analysis. Thus, the ambiguity class for \textit{dit} contains the tags \textit{VERB-SG-P3} and \textit{PAP-SG} (incidentally \textit{dit} is also a noun or an adjective, so that the complete ambiguity class for \textit{dit} is \{\textit{ADJ-SG NOUN-SG PAP-SG VERB-SG-P3}\}.

Morphological information about mood is collapsed in the same way, so that a large class of ambiguity between present indicative and present subjunctive is not resolved: again this is motivated by the fact that the mood is determined by remote elements, such as, among others, connectors that can be located at (theoretically) any distance from the verb, and in practice outside the short context considered by the tagger. For instance, a concessive subordinate conjunction like \textit{bien qu'en principe ce genre de phénomènes soit fréquent.} (Though in principle this type of phenomena is [subjunctive] frequent.)

The polarity of the main verb to which a subordinate clause is attached also plays a role in the mood. For instance, compare:

\begin{itemize}
  \item Je pense que les petits enfants \textit{font} de jolis dessins. (I think that small kids \textit{make} [indicative] nice drawings.)
  \item Je ne pense pas que les petits enfants \textit{fassent} de jolis dessins. (I do not think that small kids \textit{make} [subjunctive] nice drawings.)
\end{itemize}

Consequently, verbs of the so-called first group like \textit{chanter} (to sing) are ambiguous with respect to the indicative or subjunctive mood, and forms like \textit{chantent} are tagged as \textit{VERB-P3PL} regardless of the mood. In the case of \textit{faire} (to do, to make) however, the mood information can be easily recovered as the third person plural are \textit{font} and \textit{fassent} for indicative and subjunctive respectively.

\subsection{Person}

The person seems to be problematic for a statistical tagger (but not for a constraint-based tagger). For instance, the verb \textit{pense}, ambiguous between the first- and third-person, in the sentence \textit{Je ne le pense pas} (I do not think so) is disambiguated wrong because the statistical tagger fails to see the first-person pronoun \textit{je} and selects more common third-person reading for the verb.

We made a choice to collapse together the first- and second-person verbs but not the third person. The reason why we cannot collapse also the third person is that a we have an ambiguity class that contains adjective and first- or second-person verb. In a sentence like \textit{Le secteur matières (NOUN-PL) plastiques (ADJ-PL/NOUN-PL/VERB-P1P2)} ... verb reading for \textit{plastiques} is impossible. Because noun — third-person sequence is relatively common, collapsing also the third person would cause trouble in parsing.

Because we use the same tag for first- and second-person verbs, the first- and second-person pronouns are also collapsed together to keep the system consistent. Determining the person after the analysis is also quite straightforward: the personal pronouns are not ambiguous, and the verb form, if it is ambiguous, can be recovered from its subject pronoun.

\subsection{Lexical word-form}

We also collapsed under a same lexical item surface forms that may be attached to different lemmata (lexical forms) while sharing the same categories, such as \textit{peigner} derived from the verb \textit{peigner} (to comb) or \textit{peindre} (to paint). However, in the case of \textit{suis} first person singular of the auxiliary \textit{être} (to be) or of the verb \textit{suivre} (to follow), the distinction is maintained, as we introduced special tags for auxiliary (this distinction allows the tagger to take into account specific distributional properties of auxiliaries).
2.2.4 Gender and number

We have not introduced in our tagset gender distinctions as far as nouns and adjectives (and incidentally determiners) are concerned. Thus a feminine noun like chaise (chair) and a masculine noun like tabouret (stool) both receive the same tag NOUN-SG. Similarly masculine and feminine adjectives receive the same tag.

However, we have introduced distinctions between singular noun (NOUN-SG), plural nouns (NOUN-PL) and number-invariant nouns (NOUN-INV) such as taux (rate/rates). Similar distinctions apply to adjectives and determiners. The main reason for this choice is that number, unlike gender, plays a major role in French with respect to subject/verb agreement.

The Noun/Verb ambiguity being one of the major cases that we want the tagger to resolve, we certainly want to keep the number distinction. In the mean time, the gender distinction is of less value, not considering the complications that such a distinction may add: for instance, we would need to consider classes for masculine, feminine and gender-invariant nouns, combined with the number distinctions described above, which would make nine categories for nouns and nine for adjectives. This would require the tagger to collect enough information on sequences of such tags, which may in turn be more difficult than with our more restricted choice, especially as some combinations of tags, such as ADJ-INV followed by NOUN-INV may be rare in a training corpus.

Also, previous experiences with broad coverage parsing of possibly erroneous texts have shown that gender agreement is not as essential as on would think when it comes to parse French. This choice is also supported by our tagging experience, as only very few tagging errors involve gender agreement. Actually, on a test suite (Chanod and Tapanainen, 1995) extracted from the newspaper Le Monde (12 000 words) we counted only two errors that violated gender agreement (using the constraint-based tagger). For instance, in:

L’armée interdit d’autre part le passage... (On the other hand the army forbids the passage...)

*interdit* is mistakenly tagged as an adjective rather than a finite verb, while *armée* is a feminine noun and *interdit* a masculine adjective, which makes the noun adjective sequence impossible in this particular sentence. It should be noted, however, that sequences such as feminine noun followed by masculine adjective cannot be generally ruled out, due to possible coordinations or insertions of noun complements as in un cours (Masc) de cuisine (Fem) exceptionnel (Masc) (an exceptional cooking class) where the masculine adjective *exceptionnel* agrees with the masculine noun *cours* and not with the feminine noun *cuisine*.

An interesting, albeit minor, interest of not introducing gender distinction is that there is no problem with tagging phrases like mon allusion (my allusion) where the masculine form of the possessive determiner *mon* precedes a feminine singular noun that begins with a vowel, for euphonic reasons.

2.2.5 The rest

We avoid categories that are too small, i.e. rare words that do not fit into an existing category are collapsed together. Making a distinction between two syntactic features is not useful if there is no evidence about these categories in the training sample. We made a category MISC for all those miscellaneous words that do not fit into any existing category. This accounts for words, such as: interjection oh, salutation bonjour, onomatopoeia miaou, wordparts i.e. words that only exist as part of a multi-word expression, such as *a priori*, as part of *a priori*. Those words are also usually unambiguous.

2.3 Dividing a category

In a few instances, we introduced new categories for words that have a specific syntactic distribution. For instance, we introduced a word-specific tag PREP-DE for words *de*, *des*, *d*’ and *du*, and tag PREP-A for words *à*, *au* and *aux*. Word-specific tags for other prepositions should probably be considered too. The other readings of the words were not removed, e.g. *de* is, ambiguously, still a determiner as well as PREP-DE.

3 Building the lexicon

The initial morphological analyser is based on a lexical transducer (Karttunen et al., 1992; Karttunen, 1994). The source lexicon and rules are represented as in the two-level model (Koskeniemi, 1983). They are compiled into a single finite-state transducer using Xerox lexical tools (Karttunen and Beesley, 1992; Karttunen, 1993). The transducer maps each inflected surface form of a word to its canonical lexical form followed by the appropriate morphological tags.

This general lexicon contains all the inflectional information. For instance, word *danses* (the plural of the noun *dance*) or a second person form of the verb *danser (to dance)* has the following analyses1:

---

1The tags represent: present indicative, singular,
danser +IndP +SG +P2 +Verb
danser +SubjP +SG +P2 +Verb
danse +Fem +PL +Noun

Forms that include clitics are analysed as a sequence of items separated by the symbols < or > depending on whether the clitics precede or follow the head word. For instance j’aime (I love) is analyzed as:

je +Nom +InvGen +SG +P1 +PC
< aimer +IndP +SG +P1 +Verb

Similarly, vient-il (does he come, lit. comes-he) is analyzed as:

venir +IndP +SG +P3 +Verb
> il +Nom +Masc +SG +P3 +PC

We derived from this basic morphological transducer a new lexicon specific to the tagger, i.e. a lexicon that matches the reduced tagset described above. This involved two major operations:

- handling cliticized forms appropriately for the tagger’s needs.
- switching tagsets

In order to reduce the number of tags, cliticized items (like j’aime or vient-il) are split into independent tokens for the tagging application. This splitting is performed at an early stage, before dictionary lookup, by the tokenizer. In order to keep track of the useful information that the tokens where initially agglutinated, forms like vient-il are tokenized as vient- and il and forms like chante-t-il are tokenized as chante-t- and il. This requires in turn that forms like vient- and chante-t-be introduced in the new lexicon.

This is performed by extracting from the initial lexicon cliticized forms, removing all letters following the last hyphen on the surface side, and removing all symbols following the > symbol on the lexical side (the > symbol included). Those operations can be described by simple regular expressions that are compiled as transducers and then composed with the initial lexicon. For instance, the surface form vient-il is transduced into vient- while the corresponding lexical form:

vement +IndP +SG +P3 +Verb
> il +Nom +Masc +SG +P3 +PC

is transduced as:

venir +IndP +SG +P3 +Verb

second person, verb; present subjunctive, singular, second person, verb; and feminine, plural, noun

The tags for je represent: nominative, gender invariant, singular, first person, clitic pronoun.

This creates a new two-level lexical item, that matches vient- with venir +IndP +SG +P3 +Verb.

Meanwhile all cliticized forms like vient-il and chante-t-il are eliminated from the initial lexicon by a simple composition operation that discards all items that contain the symbols < or > on the lexical side.

With respect to switching tagsets, we use a set of contextual two-level rules, that turn the initial tags to new tags or to the void symbol if old tags must simply disappear. For instance, the symbol +Verb is transformed into +VERB-P3SG if the immediate left context consists of the symbols +SG +P3. The symbols +IndP, +SG and +P3 are then transduced to 0 (the void symbol), so that vient (or even the new token vient-) gets merely analyzed as +VERB-P3SG instead of +IndP +SG +P3 +Verb.

A final transformation consists in associating a given surface form with its ambiguity class, i.e. with the alphabetically ordered sequence of all its possible tags. For instance des is associated with the ambiguity class [+NOUN-PL +VERB-P1P2], i.e. it is either a plural noun or a verb form that belong to the collapsed first or second person paradigm (actually des is second person only, but the reasons for collapsing the two persons have been described earlier).

4 Guesser

Words not found in the lexicon are analysed by a separate finite-state transducer, the guesser. We developed a simple, extremely compact and efficient guesser for French. It is based on the general assumption that neologisms and uncommon words tend to follow regular inflectional patterns.

The guesser is thus based on productive endings (like ment for adverbs, ible for adjectives, er for verbs). A given ending may of course point to various categories, e.g. er identifies not only infinitive verbs but also nouns, due to possible borrowings from English. For instance, the ambiguity class for killer is [NOUN-SG VERB-INF].

Those endings belong to the most frequent ending patterns in the lexicon, where every rare word weights as much as any frequent word. Endings are not selected according to their frequency in running texts, because highly frequent words tend to have irregular endings, as shown by adverbs like jamais, toujours, peut-être, hier, demain, souvent (never, always, maybe...).

Similarly verb neologisms belong to the most regular conjugation paradigm characterized by the infinitive ending er, e.g. déballaduriser. The same
applies to adjectives.

With respect to nouns, we first selected productive endings \textit{iste, eau, eur, rice...} as for other categories, until we realized a better choice was to assign a noun tag to all endings, with the exception of those previously assigned to other classes. In this latter case, two situations may occur: either the prefix is shared between nouns and some other category (such as \textit{ment}), or it must be barred from the list of noun endings (such as \textit{aint}, an inflectional marking of third person plural verbs). Actually, we introduced some hierarchy in the endings: e.g. \textit{ment} is shared by adverbs and nouns, while \textit{iquement} is assigned to adverbs only.

We finally split those noun endings according to number. Not having to consider gender made the work easier, as it is less trivial to decide on the gender attached to some noun endings.

Guessing based on endings offers some side advantages: unknown words often result from alternations occurring at the beginning of the word, the rest being unchanged (e.g. derivative prefixes as in \textit{urado-jordano-palestiniene} but also oral transcriptions as \textit{les z’oreilles (the ears)}, with ‘z’ marking the phonological liaison). Similarly, spelling errors which account for many of the unknown words affect less the ending than the internal structure of the word (e.g. the misspelt verb forms \textit{appellaitent, geulait}). Also hyphens used to emphasize a word (e.g. \textit{har-mo-m-ser}) leaves endings unaltered. Those side advantages do not operate however when the alternation (prefix, spelling error) applies to a frequent word that does not follow regular ending patterns. For instance, the verb form \textit{construit}, the adverb \textit{tros} respectively misspelt as \textit{construit} and \textit{très}) are not properly recognized. More generally, the guesser does not recognize words belonging to closed classes (conjunctions, prepositions, etc.) as closed classes are presumably described in the basic lexicon. A possible improvement to the guesser would be to incorporate frequent spelling errors for words that are not otherwise recognised.

The guesser is produced by simple regular expressions such:

\begin{quote}
\texttt{Letter\textsuperscript{+} m e n t +ADV}
\texttt{Letter\textsuperscript{+} [a b l e s \mid i b l e s] +ADJ-PL}
\end{quote}

where \texttt{Letter\textsuperscript{+}} represents any letter followed by the Kleene plus operator. The case of nouns is slightly different, as all sequences of letters get a noun tag, with the exception of endings exclusively assigned to other categories. The minus operator performs this exclusion quite easily with expressions like:

\begin{quote}
\texttt{Letter\textsuperscript{+} \{-[Letter\textsuperscript{+} i q u e m e n t]\}}
\end{quote}

Various numerical expressions and complex punctuation marks are also handled by the guesser.

4.1 Testing the guesser

We extracted from a corpus of newspaper articles (L\'ib\'eration), a list of 13 500 words unknown to the basic lexicon\(^1\). Of those unknown words, 9385 (i.e. about 70 \%) are capitalized words, which are correctly and unambiguously analyzed as proper nouns by the guesser with more than 95 \% accuracy. Errors are due mostly to foreign capitalized words which are not proper nouns (such as \textit{Eight}) and onomatopoeia (such as \textit{Oooh}).

The test on the remaining 4000 unknown words which are not capitalized is more interesting. We selected randomly 800 of these words\(^4\) and ran the guesser on them. 1192 tags were assigned to those 800 words by the guesser, that is an average of 1.5 tag per word.

For 113 words, at least one required tag was missing (118 tags were missing as a whole, 4 words were lacking more than one tag: they are misspelt irregular verbs that have not been recognized as such). It means that 86 \% of the words got all the needed tags from the guesser.

273 of the 1192 tags were classified as irrelevant, and they concerned 244 words, which means that 70 \% of the words did not get any irrelevant tags.

Finally 63 \% of the words got all the required tags and only those.

If we combine the evaluation on capitalized and non capitalized words, this means that 85 \% of all unknown words are perfectly tagged by the guesser, and 92 \% get all the required tags (with possibly some unwanted ones).

The test on the non-capitalized words was tough enough, as we counted as irrelevant any tag that would be morphologically acceptable on general grounds, but which is not for a specific word. For instance, the misspelt word \textit{statisticicien} is tagged as \texttt{[ ADJ-PL NOUN-PL ]}; we count the \texttt{ADJ-PL} tag as irrelevant, on the ground that the underlying correct word \textit{statisticicien} is noun only (compare with the adjective \textit{platonicien}).

The same occurs with words ending in \textit{ement} that are systematically tagged as \texttt{[ ADV NOUN-SG ]}, unless a longer ending like \textit{iquement} is recognized. This often, but not always, makes the \texttt{NOUN-SG} tag irrelevant.

\(^1\)On various large newspaper corpora we got an average of 18 \% words that are unknown: this is mostly due to the high frequency of proper nouns.

\(^4\)performing the test took almost as much time as building the guesser.
On the other end, some improvement of the guesser would be easy to implement for endings like *iques* associated with [ *ADJ-PL NOUN-PL* ], while the adjective tag is much preferred. The noun reading is however acceptable for the corresponding singular ending *ique*, as it is widely used for neologism, e.g. *domotique*, mostly used in the singular.

As for missing tags, more than half are adjective tags for words that are otherwise correctly tagged as nouns or past participles (which somehow reduces the importance of the error, as the syntactic distribution of adjectives overlaps with those of nouns and past participles). Among those missing adjectives, quite a few represent nationalities.

The remaining words that lack at least one tag include misspelt words belonging to closed classes (e.g., *trés, vare*), or to irregular verbs (*constuit*), barbarisms resulting from the omission of blanks (*propose*), or from the adjunction of superfluous blanks or hyphens (*quand-même, so cîté*). We had also a few examples of compound nouns improperly tagged as singular nouns, e.g. *rencontres-télé*, while the plural marking only appear on the first element of the compound.

Finally, foreign words represent another class of problematic words, especially if they are not nouns. We found various English examples (at, born, of, enough, easy) but also Spanish, e.g. *levantarse* and Italian, e.g. *palazzo*, ones.

References


A The restricted tag set

In this appendix the tag set is represented. Besides the following tags, there may also be some word-specific tags like *PREP-DE*, which is the preposition reading for words *de, des* and *du*, i.e. word *de* is initially ambiguous between *PREP-DE* and *PC*. This information is mainly for the statistical tagger to deal with, for instance, different prepositions in a different way. The constraint-based tagger does not need this because it has direct access to word forms anyway. After disambiguation, the word-specific tags may be cleaned. The tag *PREP-DE* is changed back into *PREP*, to reduce the redundant information.

- **DET-SG**: Singular determiner e.g. *le, la, mon, ma*. This covers masculine as well as feminine forms. Sample sentence: *Le chien dort dans la cuisine*. (The dog is sleeping in the kitchen).
• DET-PL Plural determiner e.g. les, mes. This covers masculine as well as feminine forms. Sample sentence: Les enfants jouent avec mes livres. (The children are playing with my books.)

• ADJ-INV Adjective invariant in number e.g. heureux. Sample sentence: Le chien est heureux quand les enfants sont heureux. (The dog is happy when the children are happy.)

For statistical taggers all the adjectives may be made ambiguous between the premodifying and postmodifying position in the noun phrase to carry the information about its role in the noun phrase and to have mutually different transition probabilities. Possible tags are ADJ-INV-PRE, ADJ-INV-POST, ADJ-SG-PRE, etc.

• ADJ-SG Singular adjective e.g. gentil, gentille. This covers masculine as well as feminine forms. Sample sentence: Le chien est gentil. (The dog is nice.)

• ADJ-PL Plural adjective e.g. gentils, gentilles. This covers masculine as well as feminine forms. Sample sentence: Ces chiens sont gentils. (These dogs are nice.)

• NOUN-INV Noun invariant in number e.g. souris, Français. This covers masculine as well as feminine forms. Sample sentence: Les souris dansent. (The mice are dancing.)

• NOUN-SG Singular noun e.g. chiens, fleurs. This covers masculine as well as feminine forms. Sample sentence: C'est une jolie fleur. (It is a nice flower.)

• NOUN-PL Plural noun e.g. chiens, fleurs. This covers masculine as well as feminine forms. Sample sentence: Nous aimons les fleurs. (We like flowers.)

• VAUX-INF Auxiliary verb, infinitive être, avoir. Sample sentence: Le chien vient d'être puni. (The dog has just been punished.)

• VAUX-PRP Auxiliary verb, present participle étant, ayant.

• VAUX-PAP Auxiliary verb, past participle e.g. été, eu. Sample sentence: Le théorème a été démontré. (The theorem has been proved.)

• VAUX-P1P2 Auxiliary verb, covers any 1st or 2nd person form, regardless of number, tense or mood, e.g. 1st person singular present indicative, 2nd person plural imperative: ai, soyons, es. Sample sentence: Tu es fort. (You are strong.)

• VAUX-P3SG Auxiliary verb, covers any 3rd person singular form e.g. avait, sera, es. Sample sentence: Elle est forte. (She is strong.)

• VAUX-P3PL Auxiliary verb, covers any 3rd person plural form e.g. ont, seront, avaient. Sample sentence: Elles avaient dormi. (They had slept.)

• VERB-INF Infinitive verb e.g. danser, finir, dormir. Sample sentence: Le chien aime dormir. (The dog enjoys sleeping.)

• VERB-PRP Present participle e.g. dansant, finissant, aboyant. Sample sentence: Le chien arrive en aboyant. (The dog is coming and it is barking.)

• VERB-P1P2 Any 1st or 2nd person verb form, regardless of number, tense or mood e.g. 1st person singular present indicative, 2nd pers plural imperative: chante, finissons. Sample sentence: Je chante. (I sing.)

• VERB-P3SG Any 3rd person singular verb form e.g. chantera, finit, aboie. Sample sentence: Le chien aboie. (The dog is barking.)

• VERB-P3PL Any 3rd person plural verb form e.g. chanteront, finissent, aboient. Sample sentence: Les chiens aboient. (The dogs are barking.)

• PAP-INV Past participle invariant in number e.g. surpris. Sample sentence: Le chien m'a surpris. (The dog surprised me.)

• PAP-SG Singular past participle e.g. fini, finie. This covers masculine as well as feminine forms. Sample sentence: La journée est finie. (The day is over.)

• PAP-PL Plural past participle e.g. finis, finies. This covers masculine as well as feminine forms. Sample sentence: Les travaux sont finis. (The work is finished.)

All past participles that are marked in the basic morphological lexicon as gender/number dependant are considered by the tagger lexicon as being potentially adjectives as well as past participles.

• PC Non-nominitive clitic pronoun such as me, le. Sample sentence: Il me l'a donné. (He gave it to me.)
• **PRON** 3rd person pronoun, relative pronouns excluded. e.g. il, elles, chacun. Sample sentence: *Il a parlé à chacun.* (He spoke to every person.)

• **PRON-P1P2** 1st or 2nd person pronoun e.g. je, tu, nous. Sample sentence: *Est-ce que tu viendras avec moi?* (Will you come with me?)

• **VOICILA** Reserved for words voici and voilà. Sample sentence: *Voici mon chien.* (Here is my dog.)

• **ADV** Adverbs e.g. finalement. Sample sentence: *Le jour est finalement arrivé.* (The day has finally come.)

• **NEG** Negation particle. Reserved for the word ne. Sample sentence: *Le chien ne dort pas.* (The dog is not sleeping.)

• **PREP** Preposition e.g. dans. Sample sentence: *Le chien dort dans la cuisine.* (The dog sleeps in the kitchen.)

For statistical taggers this group may be divided into subgroups for different preposition groups, like PREP-DE, PREP-A, etc.

• **CONN** Connector. This class includes coordinating conjunctions such as et, subordinate conjunctions such as lorsque, relative or interrogative pronouns such as lequel. Words like comme or que which have very special behaviour are not coded as CONN. Sample sentence: *Le chien et le chat dorment quand il pleut.* (The dog and the cat sleep when it rains.)

For statistical taggers this group may be divided into subgroups for different connectors, like CONN-ET, CONN-Q, etc.

• **COMME** Reserved for all instances of the word comme. Sample sentence: *Il joue comme un enfant.* (He plays like a child.)

• **CONJQUE** Reserved for all instances of the word que.

• **NUM** Numeral e.g. 12,7, 120/98, 34+0.7.

• **HEURE** String representing time e.g. 12h24, 12:45:00.

• **MISC** Miscellaneous words, such as: interjection oh, salutation bonjour, onomatopoeia miaou, wordparts i.e. words that only exist as part of a multi-word expression, such as priori, as part of a priori.