

## **Ambiguity rates**

### **Automatic analysis of French text corpora and computation of ambiguity rates for different tagsets**

Éric Laporte, Max Silberztein, ASSTRIL  
May 15, 1996

**Abstract:** We analysed a French textual corpus in order to evaluate its rate of lexical ambiguity (number of lexical tags per word). Since this rate theoretically depends on the tagset and on whether compounds are delimited by tagging, the experiment was repeated with eight different tagsets. The results show that, although the information content of the tags is very different depending on the tagsets, the variation of the rate of lexical ambiguity is limited: when one shifts from the least to the most informative of the tagsets, the rate increases only from 1.6 to 2.0 tags per word.

Intoduction	2
1. Structure of output of grammatical tagging	2
1.1. Number of solutions	2
1.2. Context dependency	3
2. Tagsets	4
3. Rate of ambiguity	5
4. Experiments	6
Conclusions	6
References	6

Grammatical or lexical tagging is the operation of assigning grammatical information to words of a written text for later processing. Practically all types of natural-language processing would be improved by a previous lexical tagging of a good quality. For example, text-to-speech synthesis requires distinguishing non-homophonic homographs, like *lead* (the metal or the verb). For document indexing, the words that may appear in an index must be separated from grammatical words, which implies tagging the noun *till* differently from the homographic conjunction. Beyond such anecdotal examples, the technical interest of tagging comes from the fact that it is a prerequisite to sentence parsing, a key operation for natural language processing applications: without a recognition of the syntactic structure of sentences, most applications will hardly make any decisive progress.

Lexical tagging is thus one of the most traditional problems in natural language processing. However, it has not received satisfactory solutions yet. The output of current systems is not sufficiently reliable to be exploited: the number of words with wrong tags is often about 1 out of 11 (Tzoukermann et al., 1995) to 1 out of 30 (Brill, 1995), which means that most sentences contain at least one and therefore cannot be parsed.

Existing lexical taggers differ in several parameters, including linguistic parameters, for example the degree of precision of their output. The underlying difficulty is always the same: lexical ambiguities involving grammatical information, like that of *till*. Such lexical ambiguities are usually more frequent than what one's intuition would tell. Linguists are used to recognizing at first sight the parts-of-speech of words in sentences. For example, when reading:

- (1) *Je les ai revues par la suite* (I saw them again later)

one would never have the idea that *les* might be, in another context, a determiner, and *revues* a noun (journal), or that formally *par la suite* looks very much like *par la fenêtre* (through the window). However, for a tagger, discarding such hypotheses is a technical difficulty.

Evaluating the degree of lexical ambiguity of texts requires performing a lexical tagging. In this report, we survey the main differences between the various types of lexical taggers, and we discuss about the parameters relevant to the evaluation of the degree of lexical ambiguity. Then, we relate the experiments that we performed on French texts in order to evaluate how the degree of lexical ambiguity depends on the choice of a tagset.

## 1. Structure of output of grammatical tagging

Taggers differ in the form of their output. Fundamental differences can be observed on two levels. First, in case of ambiguities, the output of the tagging may be either one solution or a variable number of solutions. Second, the tags themselves are more or less precise (we distinguish precision and accuracy). The degree of precision of the output depends on these differences.

### 1.1. Number of solutions

When the output of a tagger consists of a single solution, and when that solution is the only correct one, then the tagging is totally successful. In fact, no existing system is sure to achieve such a result: there are always sentences for which they either produce several solutions or choose a wrong one. However, some taggers produce one or several solutions for a given input sentence; others systematically produce a single solution.

Actually ambiguous sentences, i.e. sentences for which the speakers of the language find several interpretations, are seldom but some exist. Some of them have a grammatical ambiguity: then, a single-solution tagger cannot take account of this ambiguity in its output. For example, the following ambiguous sentence is built from an attested sentence:

*La municipalité a fait de nos concitoyens les plus démunis des assurés sociaux*

(The town turned the most destitute of our fellow-

citizens into social-security insured people;

The town turned our fellow-citizens into the most destitute of social-security insured people)

In the first interpretation, *des* is an indefinite determiner; in the second, it is the contraction of the preposition *de* with the definite determiner *les*. A single-solution tagger must choose one of the interpretations, although both choices are linguistically questionable.

### 1.2. Context dependency

Among several-solution taggers, one can distinguish two categories according to whether they take into account the way different ambiguous elements combine with one another. There are always many ambiguous elements in a text. If several solutions are retained for one of them, are all these solutions compatible with all of the solutions retained for a neighbouring ambiguous element? Consider for instance the following sentence, which was built from an attested sentence too:

(2) *Les socialistes ont imité la superbe gaulliste*

(The socialists imitated the gaullists' arrogance;

The socialists imitated the beautiful gaullist woman)

Both words *superbe* and *gaulliste* are ambiguous between noun and adjective. However, in any interpretation of the sentence, one can exclude the possibility that both are nouns or both are adjectives. In other words, not all of the different tags for *superbe* are compatible with all tags for *gaulliste*. In such a situation where two ambiguities are combined, we will use the term tagging to denote a sequence of lexical tags, one for each word in a sentence. When the output of a several-solution tagger takes the form of taggings, it can take into account the fact that not all of the mathematically possible combinations may necessarily be retained. Thus, for (2), the following taggings may be retained:

(...) *la,Det superbe,N gaulliste,Adj* + (...) *la,Det superbe,Adj gaulliste,N*

and the following rejected:

(...) *la,Det superbe,N gaulliste,N* + (...) *la,Det superbe,Adj gaulliste,Adj* +

(...) *la,Pro superbe,N gaulliste,Adj* + (...) *la,Pro superbe,Adj gaulliste,N* +

(...) *la,Pro superbe,N gaulliste,N* + (...) *la,Pro superbe,Adj gaulliste,Adj* + ...

There is no standard term for this type of taggers. They could be called context-dependent taggers.

When, on the contrary, the output consists of a list of lexical tags attached to each word, the ambiguities are considered independent of one another:

(...) (*la,Det* + *la,Pro* + *la,N*) (*superbe,N* + *superbe,Adj*) (*gaulliste,N* + *gaulliste,Adj*)

In that case we will say that the tagger is context-independent. This solution is more easily implemented, but obviously less informative. In particular, it makes it impossible to take into account the ambiguities involving the limits of compound words, for example in:

*Il s'est prononcé sur tout au moins deux fois, exemple par exemple*

(He gave a decision on everything at least twice,

example after example)

In this sentence, one can a priori recognize three compounds: *tout au moins* (at least), *au moins* (at least) and *par exemple* (for example), only one of which actually appears in the sentence: *au moins*. If a list of lexical tags is associated to each simple word independently of the ambiguities in its context, it is difficult to express that *tout au moins* and *au moins* are mutually exclusive in this sentence. On the contrary, if the output takes the form of taggings,

those can be as follows:

(...) *de, Prep tout, Pro au, PrepDet moins, Adv deux, Det +*

(...) *de, Prep tout, Pro au/moins, Adv deux, Det +*

(...) *de, Prep tout/au/moins, Adv deux, Det + ...*

The various taggings of a sentence usually have rather wide common parts. In order to avoid duplicating it, an efficient and general solution for implementing them makes use of acyclic finite automata. This solution is usually adopted by authors of context-dependent taggers (Silberztein, 1989, 1994; Koskenniemi, 1990; Rimon, Herz, 1991; Roche, 1992), though they use a number of terms to refer to finite automata: directed acyclic graph, directed acyclic word graph, sentence graph, finite-state machine, finite-state network, word lattice...

## 2. Tagsets

There is not a standard grammatical tagset for a given language. Various grammatical tagsets with a more or less informative content can be designed for a language. One can classify them according to their level of precision. The simplest tagset comprises only parts of speech, i.e. about fifteen distinct tags:

(3) *Je, Ppv les, Ppv ai, V revues, V par, Prep la, Det suite, N*

More informative, and thus more precise tagsets are obtained by adding to parts-of-speech some more information, which may be taken among the following:

- limits of those compounds which occur as continuous sequences (Laporte, 1988). For example, in (1), the lexical tagging may delimit the frozen adverb *par la suite* (later) by assigning to it a single lexical tag instead of three.

- inflectional features: gender, number... when they are relevant.

- lemma or canonical form, since words in texts are inflected forms.

One or more of these types of information may be incorporated in lexical tags. The main possibilities are summarized below, with an estimate of the number of tags in the case of French. Each one is exemplified by the most correct possible tagging of (1) within the considered tagset.

- Part of speech and delimitation of compounds (about 15 tags):

(4) *Je, Ppv les, Ppv ai, V revues, V par/la/suite, Adv*

- Part of speech and inflectional features: about 85 tags. Inflectional features are coded with the conventions of INTEX (Silberztein, 1994):

(5) *Je, Ppv:1s les, Ppv:3fp ai, V:P1s revues, V:Kfp par, Prep la, Det:fs suite, N:fs*

The tagsets of the most reputable taggers of English (DeRose, 1988; Brill, 1992, 1995), those of the Brown Corpus (97 tags; Garside et al., 1987) and the Penn Treebank (36 tags; Marcus et al., 1993) belong to this type.

- Part of speech and lemma, accounting for an order of magnitude of 100,000 tags:

(6) *Je, je. Ppv les, le. Ppv ai, avoir. V revues, revoir. V par, par. Prep la, le. Det suite, suite. N*

- Part of speech, delimitation of compounds and inflectional features, i.e. about 85 tags:

(7) *Je, Ppv:1s les, Ppv:3fp ai, V:P1s revues, V:Kfp par/la/suite, Adv*

- Part of speech, delimitation of compounds and lemma, about 200,000 tags:

(8) *Je, je. Ppv les, le. Ppv ai, avoir. V revues, revoir. V par/la/suite, par/la/suite. Adv*

- Part of speech, lemma and inflectional features, i.e. about 900,000 tags:

(9) *Je, je. Ppv:1s les, le. Ppv:3fp ai, avoir. V:P1s revues, revoir. V:Kfp par, par. Prep la, le. Det:fs suite, suite. N:fs*

When a French text is tagged with a tagset of this type, about 65 % of simple words are

ambiguous.

- Part of speech, lemma, inflectional features and delimitation of compounds, i.e. about 1,100,000 tags:

(10) *Je.je.Ppv:1s les,le.Ppv:3fp ai,avoir.V:P1s revues,revoir.V:Kfp par/la/suite,par/la/suite.Adv*

The tagsets of types (6), (8), (9) and (10) are of a much larger size than others. They are exactly those tagsets that comprise lemma among lexical information. Their large size is accounted for by the number of distinct canonical forms: several dozens of thousands, in French as in other languages. Large-sized tagsets may raise a technical problem. If they do, they can be reduced without any loss of information by replacing, in each tag, the lemma with the information required to reconstruct it from the inflected form. In French, this information consists of:

- the length of the ending to be removed from the inflected form, and
- the ending to be substituted for it.

Then, example (8) becomes:

(8) *Je,0.Ppv les,1.Ppv ai,1voir.V revues,3oir.V par/la/suite,0/0/0.Adv*

With this operation, the size of a tagset like (9) decreases from 900,000 to 1,000 tags, without any loss of information (Revuz, 1991).

A tagset like (10) comprises elaborate linguistic information. In fact, tagging is meant for giving access to all the lexical information required by applications: syntactic, phonetic or other information. Lexical tags can thus be viewed as identifiers of the lexical entries to be attached to words in texts, taking into account sense distinctions. In this view, tagging a text amounts to making an inventory of its lexical ambiguities, and to disambiguating those that can be without an elaborate analysis.

### 3. Rate of ambiguity

Choosing a tagset implies deciding which lexical ambiguities will be taken into account in tagging and which will not. The more precise the tags, the more complete the inventory of lexical ambiguities in tagged texts. Therefore, the degree of lexical ambiguity depends on the tagset.

There are several means of evaluating the degree of lexical ambiguity of tagged texts. The simplest method consists of counting the number of tags attached to each word in the text, and computing the arithmetic mean on the text. This method is quite fit for text that has been tagged by a context-independent tagger: basic objects are tagged words. However, this method of evaluation is not applicable to the output of context-dependent taggers: in particular, it provides no possibility of taking into account ambiguities regarding the delimitation of compounds (cf. section 1.2.).

If one shifts from the level of words to the level of sentences, basic objects are no longer tagged words but taggings. This is a more general framework: the output of a context-independent tagger can be put into the form of taggings, whereas the output of a context-dependent tagger cannot take the form of a list of tags attached to each simple word. This is why we resorted to a method of evaluation of lexical ambiguity for which basic objects are taggings. The output of the evaluation can be used in comparisons between two taggers of any type.

The method of evaluation is simple. For each sentence, one counts

- the length  $n$  of the sentence, i.e. the number of simple words,
- the number  $a$  of distinct taggings of the sentence.

Because of lexical ambiguities,  $a$  depends on  $n$  roughly exponentially. If we imagine a text where each simple word can be tagged in exactly two different ways, we have  $a = 2^n$ , i.e.

$$\log a = n \log 2$$

and the rate of ambiguity  $r$  is given by the following formula:

$$(11) \log r = (\log a) / n$$

Of course, lexical ambiguity is not so regular. Formula (11) gives only a good approximation of the average rate of lexical ambiguity per word in a sentence. In a text, if

$$\log r_i = (\log a_i) / n_i$$

is the value of  $\log r$  for the  $i$ -th sentence, then a value of  $\log r$  for the text can be computed as the arithmetic mean of the values for the sentences:

$$(12) \log r = (\sum n_i \log r_i) / \sum n_i = (\sum \log a_i) / \sum n_i$$

where  $\sum n_i$  is the length of the text in words. This formula yields a value of  $r$  in tags per word.

#### 4. Experiments

Experiments were performed with 8 different tagsets on an excerpt of a French novel, a text of 134 Ko with 23,600 occurrences of words. The 8 tagsets were built from the LADL's lexicon DELAF. The text was tagged by INTEX with each of the tagsets. The taggings of each sentence were counted by a program in C language in the acyclic automata built by INTEX. The rate of lexical ambiguity of each tagged text was then computed with formula (12).

The results are shown in the table below. The type of each tagset refers to one of the examples (3) to (10) above. The columns with + and - define the information content of the tagsets. The value of  $r$  is in tags per word:

Type of tagset	Part of speech	Delimitation of compounds	Inflectional features	Lemma	$r$
(3)	+	-	-	-	1.60
(4)	+	+	-	-	1.63
(5)	+	-	+	-	1.92
(6)	+	-	-	+	1.64
(7)	+	+	+	-	1.957
(8)	+	+	-	+	1.67
(9)	+	-	+	+	1.958
(10)	+	+	+	+	1.99

#### Conclusions

A more complete account of lexical ambiguity allows for a finer contextual processing and provides a better description of linguistic reality. The price to be paid for these advantages is a larger number of tags per word. However, our statistics on French text show that if one uses a dictionary in order to assign tags, the increase of the number of tags per word is very limited: the rate of lexical ambiguity increases from 1.6 to 2.0 tags per word when one shifts from a set of 15 tags to a set of 1,000 tags taking account the delimitation of compounds. We conclude that the price is worth paying...

#### References

- Brill, Eric. 1992. A simple rule-based part-of-speech tagger, in *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento (Italy).  
 Brill, Eric. 1995. Transformation-based error-driven learning and natural language

- processing: a case study in part-of-speech tagging, in *Computational Linguistics* 21(4), pp. 543-565.
- DeRose, Steven J. 1988. Grammatical category disambiguation by statistical optimization, in *Computational Linguistics* 14(1), pp. 31-39.
- Garside, Roger, Geoffrey Leech, Geoffrey Sampson. 1987. *The Computational Analysis of English*, London: Longman.
- Koskenniemi, Kimmo. 1990. Finite-state parsing and disambiguation, in *Proceedings of COLING 90*, University of Helsinki, pp. 229-232.
- Laporte, Éric. 1988. La reconnaissance des expressions figées lors de l'analyse automatique, in *Langages 90, Les expressions figées*, Paris: Larousse, pp. 117-126.
- Marcus, Mitchell, Beatrice Santorini, Maryan Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank, in *Computational Linguistics* 19(2).
- Revuz, Dominique. 1991. *Dictionnaires et lexiques. Méthodes et algorithmes*, PhD thesis, University of Paris 7, CERIL, 103 p.
- Rimon, Mori, Jacky Herz. 1991. The recognition capacity of local syntactic constraints, in *Proceedings of the Fifth Conference of the European Chapter of the ACL*, Berlin, pp. 155-160.
- Roche, Emmanuel. 1992. Text disambiguation by finite-state automata, an algorithm and experiments on corpora, in *Proceedings of COLING 92*, Nantes.
- Silberztein, Max D. 1989. *Dictionnaires électroniques et reconnaissance lexicale automatique*, PhD thesis, University of Paris 7, LADL, 176 p.
- Silberztein, Max D. 1994. INTEX: a corpus processing system, in *Proceedings of COLING 94*, Kyoto.
- Tzoukermann, Evelyne, Dragomir R.Radev, William A.Gale. 1995. Combining linguistic knowledge and statistical learning in French part-of-speech tagging, in *From Text to Tags: Issues in Multilingual Language Analysis. Proceedings of the Workshop*, Dublin.