Comparative study of oral and written French automatically tagged with morpho-syntactic information

Véronique GENDNER

LATTICE, Université Paris VII & LIMSI/CNRS
Bâtiment 508, 91403 Orsay cedex B.P. 133 FRANCE
Tél.: ++33 (0)1 69 85 80 06 - Fax: ++33 (0)1 69 85 80 88
Mél: gendner@limsi.fr - http://www.limsi.fr/TLP/gendner

Abstract

In this paper, we investigate automatic tagging of French corpora and compare morpho-syntactic properties of spoken and written language on corpora from different sources. Morpho-syntactic properties are first described according to the distribution of the 8 main POS in five corpora of about 1 million words each. The automatic tagging was made with about a hundred tags and we will describe the distinctions they allow and the reason why they were chosen. We will further discuss variation of the distinction common / proper noun and some distinctions made on the verb category . For this comparison, corpora of about 40 million words were used. These larger corpora have also been used to study the influence of corpus size on vocabularies. Our study on French shows that sources in the news domain have about 36% of noun-like items (nouns and pronouns). This strongly correlates with Hudson's earlier studies on the English Brown and LOB corpora. A task-specific dialog corpus shows the highest proportions of 43% of noun-like items. Spoken news shows about 5% less nouns and 5% more pronouns than written news.

1. Introduction

In this paper, we investigate automatic tagging of French spoken corpora. Morpho-syntactic tagging of written French has been extensively studied and evaluated (for example Vergne & Giguet 1998, Adda & al 1999,...). Much less effort has been spent so far on oral language tagging (for French see Valli & Veronis 1999).

In this study different types of French corpora have been automatically tagged. We investigate the morpho-syntactic tag distributions for these different corpora. Different tag sets are described : a reduced tag set including the main POS and a second tag set including distinctions which may be helpful for automatic speech transcription. Indeed, a large portion of automatic transcription errors in French are due to morphological homophones (*allé* gone / *aller* to go). Our long term goal is to use large corpora of written and spoken language and to estimate statistical language models including relevant morpho-syntactic information for a speech recognition system.

In this paper we describe, the distributions of the main POS (Part Of Speech) on different written and oral corpora The morpho-syntactic distributions are described both in terms of occurrences in corpus and proportions in the vocabularies. The impact of corpus size on vocabulary is measured for the different POS.

We will also show how our results correlate with those of Hudson (1994) on English.

After a presentation of the used corpora (section 2), the tag set (section 3) and the tagging procedure (section 4), we will present and analyze the results in section 5.

2. Corpora

For this study, we selected corpora used at LIMSI for speech recognition research in French. These corpora mainly concern the news domain including written news corpora and transcribed, oral, news-related shows from radio and TV. Some data arise from transcribed man-machine dialogs of a train-information corpus. Part of this study is limited to roughly one million words per source because two of the used corpora were not available in such a large quantity. Some measures are carried out for growing sizes to up to 40M words per source. For each result presented below, we will specify the size of used corpora.

In the following we list the sources of oral and written language used:

Oral language:

- man-machine dialogue transcriptions of task-oriented spontaneous speech about train travelling in France (marked as **Dial** on the graphs)[1]

- Radio and television broadcast transcriptions: they concern a larger domain that is mainly the news and include mostly prepared speech.
  * precise transcriptions[2] (**TVprecis**)
  * approximate transcriptions[3] (**TVapprox**)

Precise transcriptions include all phenomena observed in the acoustic signal, explicitly indicating hesitations, repairs, rewordings, breath noise ... These have not been written down in approximate transcriptions.

Written language:

- written newspapers *Le Monde* (**LM**)
- news dispatches from the Agence France Presse. (**afp**)

Oral corpora differ from written text in several ways. Most obvious differences include

- filler words and repetitions:

Ex: *alors je crois que {FillerWord} on met beaucoup l' accent sur les considérations de de politique intérieure*

- repairs and truncated words:

[1] Those data come from the LE-3 project 4223 ARISE

[2] Precise transcriptions were acquired from the European Project OLIVE@@.

[3] The approximate transcripts have been obtained within the European project IST-1999-10354-ALERT.

Ex: *je voudrais les horaires Paris Lyon demain matin vers trois heu<res> vers neuf heures*

- no punctuation markers.

For a better use by the tagger, breath noise and filler words were replaced with punctuations. Truncated words were suppressed from the transcriptions (this only applies to the 75k words corpus of man-machine dialogue transcription **Dial**). Repetition and repair were left as is.

Here is an example of precise transcriptions: note the repetition and the dots replacing a breath noise mark

```
... jusqu'à présent on ne révèle pas de de quelle
expérience il s'agissait, ... mais des
témoignages issus des rangs de l'opposition
irakienne réfugiée à l'étranger, ... ces
témoignages font état de contamination par
l'anthrax.
```

Compare with an approximate transcription:

```
globalement , notre sentiment c' est que si les
syndicats signent ça c' est la fin du
syndicalisme confédéré en France , ça veut dire
que les syndicats n' auront plus aucune
crédibilité comme défendant les intérêts des
travailleuses et des travailleurs en attente, ça
veut dire que ça sera une trahison
```

Tag set

The tag sets used for written language tagging vary from several tens to several hundreds of tags depending on linguistic choices and the ultimate use of the tagged corpora.

To determine our set of tags, the first criterion was to stick to something robust and easily computable on large corpora. In a first step, we distinguish the nine common Parts Of Speech (Noun, Verb, Adjective, Adverb, Pronoun, Conjunction, Determiner, Preposition, Interjection) plus a Punctuation tag.

Since the morpho-syntactic tags are meant to supply the recognizer with relevant information in particular to disambiguate homophones, it seems necessary to distinguish information of number and gender. An analysis of speech recognition errors in favorable conditions (read speech) shows that one error in four is a confusion between two morphological homophones, including male / female confusion (eg: *médiatisé* / *médiatisée*) or singular / plural confusion (eg: *illicite* / *illicites*), but also tense or mode confusion (eg: *encaisser* / *encaissé*, *chantez* / *chanté*). In light of this, we decided to add mode and tense distinction to the tag set but not distinction of person which does not appeared relevant according to this analysis of speech recognition errors.

We are now using a set of 99 tags. Fig. 1 gives the detail of the distinctions made for each POS.

Main POS	**Distinctions**	**# of resulting tags**
Noun	Sub-cat: proper, common Gender: masc, fem, ? Number: sung, plur, ?	18
Pronoun Determiner Adjective	Gender: masc, fem, ? Number: sung, plur, ?	9 9 9
Verb	Sub-cat: aux., princ Mode: infinitive, indicative, subjunctive, conditional, imperative, participlepresent, preterit, past, future, past participle, don't apply Gender: masc, fem, ? Number: sing, plur, ?	40
Adverb Conjunction Preposition Interjection	-	1 1 1 1
Numeral	sing, plur, fem sing, masc sing, indet.	5
Punctuation	opening, closing, strong, weak, unspec	5
	TOTAL	**99**

Fig. 1 Distinctions in tag set

4. Tagging

Different taggers were tested on small representative parts of the corpora and similar results could be observed. To process the huge amounts of data needed for our final purpose (Language Model training),. the batch version of the tagger included in the Cordial orthographic corrector[4] appeared to be particularly robust, fast and easily available.

One of the other taggers tested was the Brill tagger with the parameters trained on French at l'INaLF, to participate in the GRACE evaluation campaign (Adda & al 1999). Although it does not provide as much information as Cordial does (no distinction of gender, for example), the number is often produced by the Brill tagger when Cordial provides an unspecified output.

Example:

Cordial:

le	serf	des	temps	féodaux	avait	plus	de	chance
D:ms	N:ms:c	D:.p	**N:m.:c**	A:mp	V:-s:mii	ADV	PREP	N:fs:c

Brill:

le	serf	des	temps	féodaux	avait	plus	de	chance
D:ms	N:ms:c	D:.p	**N:mp:c**	A:mp	V:-s:mii	ADV	PREP	N:fs:c

Corrected version:

le	serf	des	temps	féodaux	avait	plus	de	chance
D:ms	N:ms:c	D:.p	**N:mp:c**	A:mp	V:-s:mii	ADV	PREP	N:fs:c

Number specification has been added for about 4% of the running text. Unfortunately the correction process is not optimized and hence very time consuming. If the information proves useful for our application in speech recognition, a more efficient procedure has to be implemented in order to process all available data.

It is important to note that the tagging accuracy is closely related to tokenisation issues. All mentioned

[4] the spelling checker Cordial is a product of Synapse Développement, Toulouse, France.

corpora were available in a format suitable for speech recognition, which means tokenization is done to maximize lexical coverage: numbers are written out as words, rarely observed acronyms are sequences of separate letters, ... An appropriate preprocessing is necessary to adapt tokenisation for tagging because the most suitable tokenisation for speech recognition is not always suitable for tagging: for example numbers are better tagged if not written out as words.

In order to emphasize the differences of distributions of the main POS, in the results presented below, punctuations, numerals and a few other forms have not been taken into account. All of them represent less than 1% in the vocabulary. In written text, punctuations cover about 14% of occurrences. Punctuation are introduced in oral transcriptions both by transcribers, in order to facilitate comprehension and by converting acoustic marks such as breath noise or filler words in order to help automatic tagging. Therefore, approximate transcription contain only about 10 % of punctuations introduced by the transcribers whereas precise transcription contain over 15% punctuation due to the numerous acoustic marks. We have to underline that there is no direct correspondence between pauses and punctuations. In many cases, where the listener's focus is required, breath pauses are taken within phrases, before an important information. It also frequently happens that there is no pause to introduce new phrase or sentences. Since punctuation represent very different phenomena in oral and written corpora, we decided to exclude them from comparisons.

In many cases, it is difficult to determine whether a numeral is a adjective, a noun or a pronoun and the tagger's errors even add to this ambiguity. Therefore, we decided to identify numerals with a specific tag and to ignore them in the following results.

The tagging of interjection did not seem very accurate, it was therefore ignored as well as some ill formed tags resulting from bugs in the tagger's output. This mixed class of ignored forms represents less than 0,5% of occurrences. Fig. 2 shows the weight of each of the three ignored categories both in the vocabularies and in the corpora.

	Punctuations	Numerals	Other (errors, interjection, euphonic part. …)
Entries in vocabularies	< 0,1 %	± 0,3 %	± 1,0 %
Occurrences in corpora	10-15 %	1-2 %	< 0,5 %

Fig. 2 Forms ignored in the following results.

5. Analysis & results

A comparison of the distributions of the main POS in corpora of about 1 M words are shown on Fig. 3. A general observation concerns the proportion of nouns which is highest (between 25% and 32%) for all considered corpora. The first represented corpus (**Dial**) in Fig.3 corresponds to the train information dialogues where typical sentences are:

je	voudrais	un	train	de	Lyon	à	Marseille
Pron	V	Det	N	Prep	N	Prep	N
I'd	*like*	*a*	*train*	*from*	*L.*	*to*	*M.*

je	veux	aller	de	Paris	à	Bordeaux
Pron	V	V	Prep	N	Prep	N
I	*want*	*to go*	*from*	*P.*	*to*	*B.*

These example sentences contain respectively 3/2 nouns, 2/2 prepositions, 1/2 verbs, 1/1 pronouns, 1/0 determiners. This may contribute to explains the figure of Det (13%) for Dial which is significantly low whereas the proportion of verbs (19%) is particularly high.

For [**Dial**] the main POS ranked by decreasing frequency in the corpus are:

N (32%), V (19%), Prep (18%), Det (13%), Pron (11%), Conj (3%), Adj (2%), Adv (2%).

For all other investigated corpora (oral and written news) the determiner is the second most frequent POS with more than 16% and conjunction the less frequent POS.

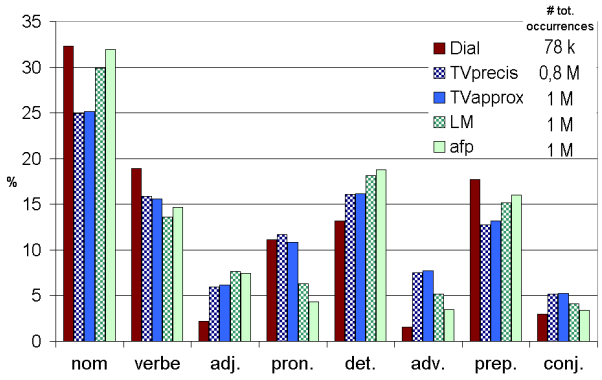

Fig. 3 Proportions of the main POS occurring in corpora for different types of data (excluding punctuation, interjection and numerals from the tagged corpora).

The 2 distributions for the oral transcribed corpora (**TVprecis** and **TVapprox**) are very close. The differences in the proportions for precise and approximate transcriptions of radio-TV broadcast do not exceed 2 points for any POS. The same is true for the two written corpora ([**LM**] and [**afp**]). However, precise and approximate transcriptions of radio/TV broadcasts differ significantly from written newspapers by having a lower proportion of nouns (25% vs 31%) and a much higher proportion of pronouns (11% vs 6%).

The particularly high proportion of nouns (over 30%) observed for the news dispatches is likely to be due to the very factual style of such data (on contrary, it shows only a few percents of adverbs). For a discussion on the correlation between the proportions of nouns and the informational character of a text see Hudson (1994).

As one could expect, the smaller proportions of nouns in the oral corpora are accompanied by also smaller figures for determiners and prepositions as compared to the written news corpora. In oral corpora the absolute figures for prepositions are slightly lower than those for verbs. This tendency is reversed in written data.

It is interesting to note that our figures for nouns and pronouns correlates with Hudson's observation on English that about 37% of occurrences in any type of corpora is nominal, that is nouns or pronouns. For example, in a study done on two larger corpora of 40 million words each, we found 36% nominals, for both written newspaper and oral transcriptions. Those represent 24% of common nouns, 6% of proper nouns and 6% pronouns in written newspapers and respectively 21%, 4% and 11% in approximate transcriptions. That is to say nouns are more frequent in written newspapers, essentially due to more occurrences of proper nouns: in the written newspapers common nouns cover 5,5 times more occurrences than proper nouns, whereas in oral transcription, this factor is only of 4. In other words, proper nouns are less frequent in oral corpora and the weight of common noun is higher.

One must keep in mind that punctuations, numerals and a few other diverse forms (see Fig. 2) are not taken into account to measure our proportion of POS and that Hudson does not specify whether his figures include them or not.

Using the Brown corpus and the LOB corpus[5], Hudson also compares the proportion of proper and common nouns in what he calls the "informational" and "imaginative" parts of the corpora. In the "informational" parts of the two corpora, common nouns represent about 25% of occurrences, proper nouns cover 5% and pronoun 6%. This is strikingly similar to what we found for the French newspaper *Le Monde*: 24% of occurrences for common nouns and 6% for proper nouns and 6% for pronouns.

Hudson (1994) and Biber (1988) note that subcorpora above average for prepositions are also above average for common nouns and vice versa and that a similar trend links verbs and pronouns. This can also be observed on our data. Nevertheless, they indicate that every corpus which is high in prepositions and common nouns is lower on verbs and pronouns, and vice versa. For our data, this is true of the four corpora in the news domain but not for the man-machine dialogue corpus: the proportion of noun and prepositions is the highest and the proportion of verbs is also the highest. This shows again the very specific character of this corpus.

The main category "Verb" accounts for similar proportions of word tokens, regardless of the type of data (between 13 and 17%) but more precise tags reveal significant differences: the conditional mode is omnipresent in the man-machine dialogues (over 30% vs 2 % in other corpora) due to highly frequent requests introduced by (*je voudrais, j'aimerais* I would like to). Past participles and auxiliary verbs are much more frequent in news dispatches (resp. 35% & 24%) than in oral transcriptions (from 3% for both in the dialogues to 17% and 13% in broadcast transcriptions).

Fig.4 displays the proportions of the main POS in vocabularies. By vocabularies we mean the list of distinct items per corpus. The vocabulary of the **Dial** corpus (75 k words) contains only about 2k items, TV transcripts about 40k items and written news more than 50k entries (for corpora sizes of about 1M words).

We can observe that for all corpora, nouns have the largest rate of vocabulary entries (they account for at least 50%), followed by verb and adjective POS.

The remaining POS account together for less than 10%.

Comparing vocabularies for written and oral language in the news domain the proportions of nouns are higher for written language (corpora [LM] & [afp]) whereas verb rates are lower.

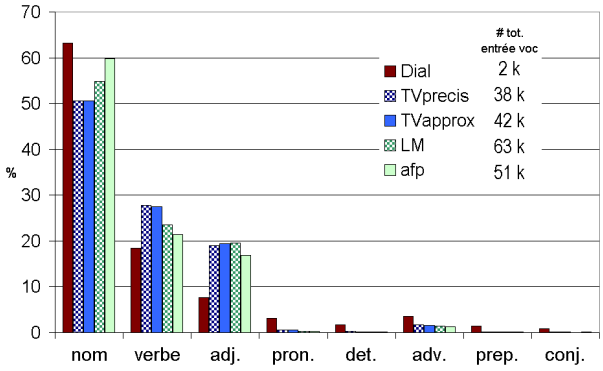

Fig 4 Proportions of the main POS in the vocabularies (Vocabularies are the lists of distinct items observed in each corpus).

The vocabulary extracted form *Le Monde* corpus (see Fig. 5) shows larger numbers for verbs, adjectives and adverbs (14000, 12000 and 800) than any other corpus (less than 10000, 8000 and 600) The numbers of verbs from written language are at least as high as the numbers from oral language. The lower proportion of verbs (Fig 4) is only due to the higher number of nouns in written corpora (32000 vs 19000).

The preponderance of proper nouns observed in occurrences for written text is also true of the entries in the vocabularies: for written texts, proper nouns account for 58% of noun entries. In the oral vocabulary they represent only 46% of the nouns. For those corpora of about 40 million words, there are about 133k proper nouns (98k common nouns) for written newspaper and 54k different proper noun (and 62k common nouns) for oral transcriptions.

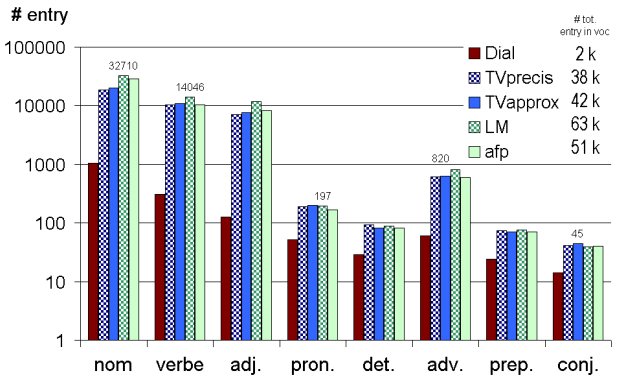

Fig. 5 Number of distinct words per POS (y-axis in log scale).

[5] The Brown Corpus of written American English was produced at Brown University and is reported in Francis & Kucera 1982. The Lancaster-Oslo-Bergen (LOB) corpus of a million words of British English is described in Johansson & Hofland 1989.

The proportions measured in Fig.4 are closely related to corpora sizes. With increasing corpus size the different POS will not keep the same proportions in vocabularies.

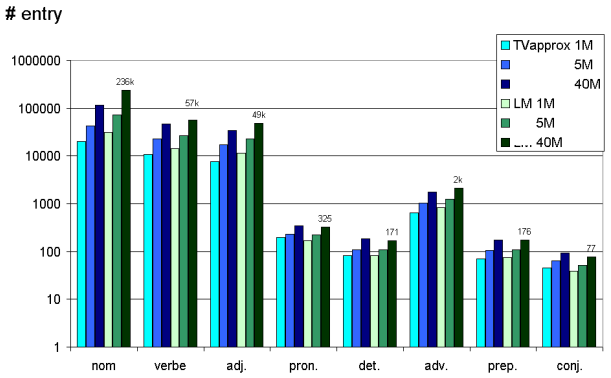

Fig. 6 Influence of corpus size on the number of distinct words per POS

If comparing corpora of different sizes (for availability reasons), it is important to keep in mind the influence of corpus size. It's effect is negligible on the proportions in corpus occurrences, but this is not true for the proportions in the vocabulary. The number of nouns increases much faster with corpus size than any other POS (Fig. 6, note the log scale). The number of verbs and adjectives also grows faster than other POS.

6. Conclusion

A first general observation is that our experiment confirms Valli & Veronis 's conclusions: contrary to what could be expected due to its specific phenomena (repairs, interruptions, filler words, etc) oral language can be easily processed with a tagger designed for the written language.

For this study on automatic tagging of French oral data compared with written data, we described the distributions of the main POS both in the vocabularies and in corpora. We also discussed some of the more precise distinctions determined according to their possible interest for speech recognition.

We have shown that some typical morpho-syntactic distinctions emerge between oral and written data: oral transcription differ form written newspapers by having a much higher proportion of pronouns (11% vs 6%) and a lower proportion of nouns (25% vs 31%). Proper nouns have a lower weight in oral corpora, both in the vocabulary and in the occurrences. The general observation Hudson describes for English that about 37% of occurrences are nouns or pronouns is also true for French news data, both oral and written. Nevertheless, with a very specific corpus such as man-machine dialogue transcriptions on a restricted domain, this general tendency does not apply anymore: a 43% rate of noun and pronoun has been measured here.

Ongoing work include POS bigram and trigram analysis and Language Model estimation on the tagged data.

7. Acknowledgement

This work is done under the supervision of Martine ADDA-DECKER whom I would like to thank very much for her many advises and remarks and for her warm support

Most of the data used in this study had previously been selected and formatted by Gilles ADDA whom I would like to thank very much for his advises about data processing among other.

8. References

Gilles Adda, Joseph Mariani, Patrick Paroubek, Martin Rajman, Josette Lecomte, *L'action GRACE d'évaluation de l'assignation de parties du discours pour le français*,

Douglas Biber, *Linguistic features: algorithms and functions* in *Variation across speech and writing*. 1988, Cambridge: Cambridge University Press.

W. Nelson Francis, Henry Kucera *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.

Richard Hudson, *About 37% of word token are nouns* Language, 1994, vol. 70:2 p331-339.

Stig Johansson, Knut Hofland, *Frequency analysis of English vocabulary and grammar, based on the LOB Corpus, I: Tag Frequencies and word frequencies.* Oxford: Clarndon Press, 1989.

Denise Malrieu, *Genres et variations morphosyntaxiques. Quelles variables pertinentes?* Journée d'étude ATALA du 28 avril 2001 http://www.atala.org/je/010428/Malrieu.pdf

André Valli, Jean Veronis, *Etiquetage grammatical des corpus de parole: problèmes et perspectives.* Revue Française de Linguistique Appliquée. 1999

Jacques Vergne, Emmanuel GIGUET, *Regard théorique sur le tagging* TALN98, p.22-31