POS-tagging for Oral Texts with CRF and Category Decomposition

Isabelle Tellier[1], Iris Eshkol[2], Samer Taalab[1], and Jean-Philippe Prost[1,3]

[1] LIFO, Université d'Orléans, France
[2] LLL, Université d'Orléans, France
[3] INRIA Lille - Nord Europe, France
{name}.{lastname}@univ-orleans.fr

Abstract. The ESLO (Enquête sociolinguistique d'Orléans, *i.e. Sociolinguistic Survey of Orléans*) campaign gathered a large oral corpus, which was later transcribed into a text format. The purpose of this work is to assign morpho-syntactic labels to each unit of this corpus. To this end, we first studied the specificities of the labels required for oral data, and their various possible levels of description. This led to a new original hierarchical structure of labels. Then, since our new set of labels was different from any of those of existing taggers, which are usually not fit for oral data, we have built a new labelling tool using a Machine Learning approach. As a starting point, we used data labelled by Cordial and corrected by hand. We used CRF (Conditional Random Fields), to try to take the best possible advantage of the linguistic knowledge used to define the set of labels. We measure accuracy between 85 and 90, depending on the parameters.

1 Introduction

Morpho-syntactic tagging is essential to text analysis, as a preliminary step to any high level processing. Different reliable taggers exist for French, but they have been designed for handling written texts and are, therefore, not suited to the specificities of less "normalised" language. Here, we are interested in the ESLO[4] corpus, which comes from records of spoken language. ESLO thus presents specific features, which are not well accounted for by standard taggers.

Several options are possible to label transcribed spoken language: one can take a tagger initially developed for written texts, providing new formal rules let us adapt it to take into account disfluences (Dister, 2007 [1]); or one can adapt the transcribed corpus to the requirements of written language (Valli and Veronis, 1999 [2]). We have chosen a different methodological approach. Starting from the output of a tagger for written language, we have first defined a new tag set, which meets our needs; then, we have annotated a reference corpus with those new tags and used it to train a Machine Learning system.

For that kind of annotation task, the state-of-the-art technology for supervised example-based Machine Learning are the Conditional Random Fields

[4] Enquête sociolinguistique d'Orléans, *i.e. Sociolinguistic Survey of Orléans*

Special issue: Natural Language Processing and its Applications.
Research in Computing Science 46, 2010, pp. 79-90

Received 27/11/09
Accepted 16/01/10
Final version 09/03/10

(CRF). CRF is a family of recently introduced statistical models (Lafferty *et al.*, 2001 [3], Sutton and McCallum, 2007 [4]), which have already proven their efficiency in many natural language engineering tasks (McCallum and Li, 2003 [5], Pinto *et al.*, 2003 [6], Altun *et al.*, 2003 [7], Sha and Pereira, 2003 [8]). Our experiments make use of CRF++[5], a free open-source library, developed by Taku Kado. We proceed with the testing of various strategies for decomposing the labels into a hierarchy of simpler sub-labels. The approach is original, in that it eases the learning process, while optimising the use of the linguistic knowledge that ruled the choice of intial labels. In that, we follow the procedure suggested by Jousse (2007 [9]), and Zidouni (2009 [10]).

In the following, Sect. 2 is dedicated to the presentation of our corpus and of the tagging process, focusing on the labelling problems raised by spoken language. After justifying the choice of a new tag set, we explain the procedure we have adopted for acquiring a sample set of correctly labelled data. In Sect. 3 we present the experiments we have carried out with CRF++, to learn a morpho-syntactic tagger from this sample. We show how the performance of the learning process can be influenced by different possible decompositions of the target labels into simpler sub-labels.

2 Oral Corpus and its Tagging

This section deals with the morpho-syntactic tagging of an oral corpus, and the difficulties that it causes to the tagger Cordial. The specificities of spoken language lead us to propose a new set of more suitable tags.

The morpho-syntactic Tagging of an Oral Corpus. The purpose of tagging is to assign to each word in a corpus a tag containing morpho-syntactic information about that word. This process can be coupled with stemming, to reduce the occurrence of a given word to its base form or *lemma*. The main difficulty of morpho-syntactic tagging is the ambiguity of words belonging to different lexical categories (*e.g.* the form *portes* in French is either a plural noun (*doors*) or the second person singular of present indicative or subjunctive of the verb *porter* (*to carry*): the tagger must assign the correct tag in a given context. Taggers usually also have problems with words which are not recognized by their dictionary: misspelled words, proper nouns, neologisms, compound words, ...

Tagging an oral corpus faces even more problems. Firstly, the transcriptions are usually not punctuated in order to avoid anticipating the interpretation (Blanche-Benveniste and Jeanjean, 1987 [11]). Punctuation marks such as comma or full stop, and casing are typographical marks. The notion of sentence, mostly graphical, was quickly abandoned by linguists interested in speech. Studies on spoken language have also identified phenomena which are specific to speech, called *disfluence*: repeats, self-corrections, truncations, *etc.* Following (Blanche-Benveniste, 2005 [12]), we believe that all these phenomena should be

[5] `http://crfpp.sourceforge.net/`

included in the analysis of language even if it raises processing issues. Elements, like *hein, bon, bien, quoi, voilà, comment dire*, (*eh, well, what, how to say, ...*) with a high frequency of occurrence in oral corpora, and without punctuation, can be ambiguous, because they can sometimes also be nouns or adjectives (as it is the case for *bon, bien* — meaning *good*). The currently existing tools for tagging are not suitable for oral, which is why this task is so difficult.

The Tagging by Cordial and its Limits. The Socio-Linguistic Survey in Orleans (Enquête Socio-Linguistique à Orléans, ESLO) is a large oral corpus of 300 hours of speech (approximately 4,500,000 words) which represents a collection of 200 interviews recorded in a professional or private context. This investigation was carried out towards the end of the Sixties by British academics in a didactic aim. The corpus is made up of 105 XML files generated by Transcriber, and converted to text format. Each file corresponds to a recording situation. Significant conventions of transcription are:

- the segmentation was made according to an intuitive unit of the type "breathing group" and was performed by a human transcriber;
- the turn-taking was defined by speaker changes;
- no punctuation except exclamation and question marks;
- no upercase letters except named entities;
- word truncation is indicated by the dash (*word-*);
- the transcription is orthographic.

The transcribed data was tagged by Cordial in order to have a reference corpus. This software was chosen for its reliability. As of today, it is one of the best taggers for French written language, with a wide range of tags, rich in linguistic information. The result of tagging is presented in a 3-column format: *word*, *lemma* and *lexical category* (POS), as exemplified in Table 1. Cordial uses about

Word	Lemma	POS
comment (*how*)	comment	ADV
vous (*you*)	vous	PPER2P
faîtes (*make/do*)	faire	VINDP2P
une (*one/a*)	un	DETIFS
omelette (*omelette*)	omelette	NCFS

Table 1. Example of tagging by Cordial.

200 tags encoding morphological information of different kinds such as gender, number or invariability for nouns and adjectives; distinction in modality, tense and person for verbs, and even the presence of aspirated 'h' at the beginning of words. However, the analysis of Cordial's outcome revealed a number of errors. The first group of errors includes "classical" errors such as the following ones.

Ambiguity: *et vous êtes pour ou contre* (*and are you for or against*) ⇒ {contre, contrer, VINDP3S} instead of {contre, contre, PREP3}.

Proper nouns: *les différences qu'il y a entre les lycées les CEG et les CES* (*the differences that exist among [types of secondary and high schools]*) ⇒ {CEG, Ceg, NPMS} instead of {CEG, CEG, NPPIG4} and {CES, ce, DETDEM} instead of {CES, CES, NPPIG}.

Locutions: *en effet* (*indeed*) ⇒ analysed in two separate lines, as opposed to a compound: {en, en, PREP}, then {effet, effet, NCMS} while it is an adverb.

We have also found errors, which are specific to oral data, as :

Truncation: by convention in ESLO, word truncation is indicated by the dash, which raises a problem for tagging by Cordial:
on fait une ou deux réclam- réclamations (*we make one or two complaints*) ⇒ {réclam- réclamations, réclamréclamations, NCMIN5} instead of analysing the sequence in two separate units: {réclam-, reclam-, NCI[6]} and {réclamations, réclamation, NCFP}

Interjection: Cordial does not recognize all the possible interjections present in oral corpora:
alors ben écoutez madame (*so uh listen madam*) ⇒ {ben, ben, NCMIN). This phenomenon also presents a problem of ambiguity since, according to Dister (2007 [1, p. 350]):

> any form can potentially become an interjection. One, then, observes a grammatical re-classification (...), the phenomenon whereby a word belongs to a grammatical class may, in speech, change.

j'ai quand même des attaches euh ben de la campagne qui est proche quoi (PRI7) (*I still have ties* [euh ben] *to the nearby countryside* [quoi])

Repeat and self-correction: *je crois que le* ({le, le, PPER3S} instead of {le, le, DETDMS}) *le* ({le, le, DETDMS}) *les saisons* (*I think that the the seasons*)

Note, as well, a number of errors such as typos or spellings made by human transcribers. The transcription was not spell-checked.

New Choice of Tags. In order to adapt the tagging to our needs, we propose a number of modifications to the tag set. Those changes are motivated on the one hand by the reduction of the number of tags without loss of the necessary linguistic information, and on the other hand, by the need to adapt the tags to spoken language and to the conventions adopted for the transcription of our corpus. We present here a (non-exhaustive) list of modifications.

- New tags were introduced, such as MI (unknown word) for cases of truncation, and PRES (announcer) for phrases such as *il y a, c'est, voilà* (*there is, this is, there it is*), both very frequent in oral.

[6] Common noun invariable in number

- A few tags, which are too detailed in Cordial, were simplified. For example, the set of tags marking the invariability of adjectives or nouns (masculine invariant in number, feminine invariant in number, singular invariant in gender, plural invariant in number and gender) were replaced by a single tag (*invariable*). The tags marking the possibility for the word to begin with an aspirated 'h' were removed.
- In order to make the system more uniform, some tags were enriched. For example, indications about the gender and number were added to the demonstrative and possessive determiners for coherence purpose with other types, such as definite and indefinite determiners.

The morpho-syntactic tags often contain information of different kinds. They always mark information about the Part-Of-Speech. That basic information seems to be the easiest to acquire from dictionary lookup except, of course, in the case of lexical ambiguity. The tags generally include additional information from different linguistic dimensions:

morphological: concerns the structure of a word as its type and number, the invariability for nouns, adjectives, pronouns and some determiners;
syntactic: describes the function of words in the sentence and they relate with each other, *e.g.* coordination and subordination for conjunctions;
semantic: related to the description of word's meaning such as feature of possessive, demonstrative, definite, indefinite or interrogative for the determiners.

In order to account for that extra information, we propose to structure the tags in 3 levels, called respectively L0 (level of POS tags), L1 (morphological level) and L2 (syntactic and semantic level). A sample of that hierarchical structure is illustrated in Fig. 1. As shown in Fig. 1, some tags:

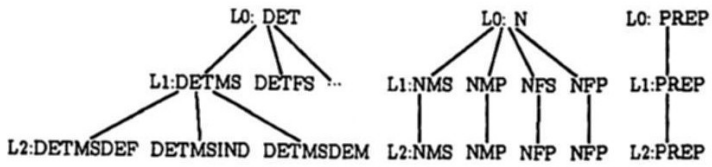

Fig. 1. Sample of the tags hierarchical structure

- remain the same on all 3 levels, *e.g.* adverbs, announcer, prepositions, ...;
- only change on level 2, such as nouns, adjectives, verbs;
- change on every level, including new information such as pronouns and determiners.

In addition to that hierarchical structure, other types of linguistic knowledge can be taken into account during tagging. According to inflectional morphology, a word is made up of a root and a sequence of letters, which often carry morphological information: in French, endings such as *-ait*, *-ais*, *-e* ,*-es* indicate the tense, gender, number, ... In inflectional morphology, those endings are called *grammatical morphemes*. When considering the root as the part shared by all

the forms of a word, it is possible to extract these final sequences from the surface form in order to determine the morphological part of the tag which must be associated with this word. That linguistic knowledge can be exploited in order to improve the performance of a Machine Learning system, as we discuss it in the next section. The reference corpus contains 18,424 words, and 1723 utterances. This data was first tagged by Cordial, and then corrected semi-automatically, in order to make it meet our new tagging conventions. The hand processing was made by linguistic students as part of a 3-month internship.

3 Experiments

We now have a reference corpus, whose labelling was manually corrected and is considered as (nearly) perfect. It is, thus, possible to use it for training a Machine Learning system. As far as learning a tagger is concerned, the best performing statistical model is the one of Conditional Random Fields (CRF) (Lafferty *et al.*, 2001 [3], Sutton and McCallum, 2007 [4]). We choose to work with it. In that section, we first briefly describe the fundamental properties of CRF, then present the experimental process, and finally we detail the results. Our goal is to maximise the use of the linguistic knowledge which guided the definition of our tag set, in order to improve the quality of the learned labels. In particular, we want to determine whether learning the full labels (*i.e.*, those containing all the hierarchical levels of information) could be improved by a sequence of intermediate learning steps involving less information. Note that we do not rely on any dictionary, which would enumerate all the possible labels for a text unit.

CRF and CRF++. CRF are a family of statistical models, which associate an observation x with an annotation y using a set of labelled training pairs (x, y). In our case, each x coincides with a sequence of words, possibly enriched with additional information (*e.g.*, if the words' lemmas are available, x becomes a sequence of pairs (word, lemma)), and y is the sequence of corresponding morpho-syntactic labels. Both x and y are decomposed into *random variables.* The dependencies among the variables Y_i are represented in an undirected graph. The probability $p(y|x)$ of an annotation y, knowing the observation x is:

$$p(y|x) = \frac{1}{Z(x)} \prod_{c \in C} \psi_c(y_c, x) \text{ with } Z(x) = \sum_y \prod_{c \in C} \psi_c(y_c, x)$$

where C is the set of cliques (*i.e.* completely connected subgraph) over the graph, y_c is the configuration taken by the set of random variables Y belonging to the clique c, and $Z(x)$ is a normalization factor. The potential functions $\psi_c(y_c, x)$ take the following form:

$$\psi_c(y_c, x) = \exp\left(\sum_k \lambda_k f_k(y_c, x, c)\right)$$

The functions f_k are called *features*, each one being weighted by a parameter λ_k. The set of features must be provided to the system, whose learning purpose is to

assign the most likely values for each λ_k according to the available valued data. Most of the time, function results are 0 or 1 (but they could also be real-valued).

In linear CRF, which are well-suited to sequence annotation, the graph simply links together the successive variables associated with the sequence elements. The maximal cliques of that kind of graph are, thus, the successive pairs (Y_i, Y_{i+1}). That model is potentially richer than the HMM one, and usually gives better results.

CRF++, the software that we are using, is based on that model. Features can be specified through *templates*, which are instanciated with example pairs (x, y) provided to the program. We kept the default templates provided by the library; they generate boolean functions using the words located within a two-word neighborhood around the current position, as exemplified in Ex. 1.

Example 1. In Table 1, the first column corresponds to the observation x, the third one to the annotation. Hence:
$x =$ "*comment vous faites une omelette*"[7],
$y =$ ADV, PPER2P, VINDP2, PPER2P, DETIFS, NCFS.
For a given position i identifying the clique $(i, i+1)$, the template tests the values of Ys in the clique, and the values of X in position $i, i-2, i-1, i+1, i+2$. At position $i = 3$ we get the following feature f:
if $Y_i =$ VINDP2 and $Y_{i+1} =$ PPER2P and $X_i =$ 'faites' and $X_{i-2} =$ 'comment' and $X_{i-1} =$ 'vous' and $X_{i+1} =$ 'vous' and $X_{i+2} =$ 'une' then $f = 1$ else $f = 0$.

The template also generates simpler functions, where only the positions i, $i-1$, and $i+1$ of X are tested, for example. With that example, we see that the disfluencies are directly taken into account in the model by the fact that they occur in the set of training examples provided to the learner.

Experimental Framework. For our experiments, the corpus was split in 10 subsets and we performed a 10-fold cross-validation. The features are mainly built from observations over words. We have also carried out experiments where the word lemma is supposed to be known. In order to enrich the data even more, we also relied on inflectional morphology, mentioned in Sect. 2:

1. the distinction between *root* and *rest*: the root is the string shared between the word and the lemma, while the rest is the difference between them; if word=lemma, by convention we note Rword=Rlemma=x, else word= Root+Rword and lemma=Root+Rlemma (where the symbol + denotes here the string concatenation);
2. the *word tail*: D_n(word) = n last letters of the word; for instance, if word= 'marchant' and lemma='marcher' then Root='march', Rword='ant', Rlemma ='er' and D_2(word)='nt'.

Reference Experiments. The reference experiments consist of learning the most detailed level (L2) directly. Six different tests were run, which we describe next. We denote by $Feat^{(\text{args})}$ the features built from (args).

Test I $Feat^{(word,\,lemma)}$: *about 10,000,000 features produced;* $F_1 = 0.86$.

Test II $Feat^{(\text{word, lemma, Rword, Rlemma})}$; *11,000,000 features;* $F_1 = 0.88$.

Test III *If* word=lemma *we use* D_2(word) *and* D_3(lemma), *hence:*
$Feat^{(\text{word, lemma, Rword}|D_2(\text{word}),\text{Rlemma}|D_3(\text{lemma}))}$; *20,000,000 feat.;* $F_1 = 0.82$.

Now, if the lemmas are unknown, we obtain the following:

Test IIIbis $Feat^{(\text{word},\,D_3(\text{word}))}$; *8,000,000 features;* $F_1 = 0.87$;

Test IV *Similar to III, but with* D_3 *everywhere:*
$Feat^{(\text{word, lemma, Rword}|D_3(\text{word}),\text{Rlemma}|D_3(\text{lemma}))}$; *20,000,000 feat.;* $F_1 = 0.89$.

Again, without relying on lemmas, we get:

Test IVbis $Feat^{(\text{word},\,D_3(\text{word}),\,D_2(\text{word}),\,D_1(\text{word}))}$; *20,000,000 feat.;* $F_1 = 0.88$.

As expected, the richer the features, the better the results. Knowing the lemmas, for instance, increases the accuracy by 2 points in average. The downside is the increased cost timewise for the learning phase, caused by a much larger number of generated features.

Cascasde Learning. In order to exploit the knowledge contained in the labels — *i.e.*, mainly their organisation in a 3-level hierarchical structure — we first learned each level independently, using the same test set (Test I to IVbis) as previously. The scores obtained are presented in Table 2. We observe that the

Level (num. of tags)	Test I	Test II	Test III	Test IV	Test IIIbis	Test IVbis
L0 (16)	0.93	0.93	0.94	0.94	0.92	0.93
L1 (72)	0.86	0.89	0.9	0.9	0.88	0.89
L2 (107)	0.86	0.88	0.82	0.89	0.87	0.88

Table 2. Accuracy measures when learning separately each of the hierarchical levels of labels.

coarser the levels in terms of how detailed the information is, the easier they are to learn. It can be explained by the reduced number of labels to be learned. Meanwhile, since each level of labels in the hierarchy depends on the previous one, one can hypothesise that using the results from the levels L_j when learning Level L_i (with$j < i$) may improve the results at L_i. The purpose of the next set of experiments is to test that hypothesis: we say that the different hierarchical levels are learned in *cascade*, as in Jousse (2007 [9]) and Zidouni (2009 [10]). In the previous set of experiments, the best scoring tests are Test III and IV; as an attempt to improve those tests, we have designed Test V and VI as follows. We denote by CRF(Li|feat(args)) the learning of level Li knowing the features based on args.

Test V *(derived from Test III)* word, lemma *and* D_3(lemma) *are used to generate the features for learning Level L0; then the result RL0 is used, with the same data, to learn Level L1; and so forth. The successive learning phases are given next.*

- $CRF(L0|feat(\text{word}, \text{lemma}, D_3(\text{lemma}))) \rightarrow ResL0$
- $CRF(L1|feat(\text{word}, \text{lemma}, D_3(\text{lemma}), ResL0)) \rightarrow ResL01$
- $CRF(L1|feat(\text{word}, \text{lemma}, D_3(\text{lemma}), ResL0, ResL01)) \rightarrow ResL012$

Test VI *(derived from TestIV) This time, the initial features are generated with* word, Rword, Rlemma, D_3(word),D_3(lemma). *The successive learning phases are the following:*

- $CRF(L0|feat(\text{word}, \text{Rword}, \text{Rlemma}, D_3(\text{word}), D_3(\text{lemma}))) \rightarrow ResL0$
- $CRF(L0|feat(\text{word}, \text{Rword}, \text{Rlemma}, D_3(\text{word}), D_3(\text{lemma}), ResL0)) \rightarrow ResL01$
- $CRF(L0|feat(\text{word}, \text{Rword}, \text{Rlemma}, D_3(\text{word}), D_3(\text{lemma}), ResL0, ResL01)) \rightarrow ResL012$

Level	III	IV	V	VI
L0	0.94	0.94	0.94	0.94
L1	0.9	0.9	0.88	0.9
L2	0.82	0.89	0.87	0.89

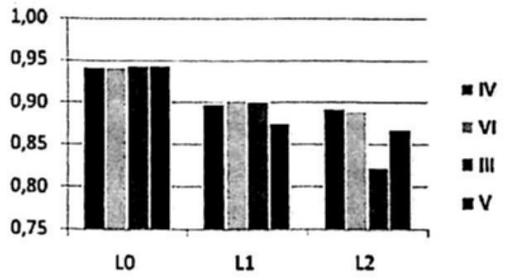

Fig. 2. Accuracy measures for Test III to VI

As shown in Fig. 2, Test V and VI give good results, but not really better than the initial Test III and IV. Therefore, unfortunately, cascade learning does not seem to improve the results obtained in the reference experiments, where L2 is learned directly. That conclusion is confirmed by the experiments without lemmas, the outcome of which we do not detail. Next, we re-consider the way the information contained in the L2 labels is decomposed, in order to better learn those labels.

Learning by Decomposition and Recombination of Labels. We decompose the L2 labels into *components*, so that they can be learned independently. We call *label component* a group of atomic symbols, which cannot all belong to the same label. Intuitively, those components correspond to the possible values for a linguistic feature, such as Gender or Number.

Example 2. The labels in $\mathcal{L} = \{\text{NFS}, \text{NFP}, \text{NMS}, \text{NMP}\}$ come from the concatenation of elements in the sets $\{N\}$ (Noun), $\{M, F\}$ (Gender), and $\{S, P\}$ (Number). All the four labels in $\mathcal{L}$ can be recombined by the cartesian product of three components: $\{N\}\cdot\{M, F\}\cdot\{S, P\}$, where $\cdot$ (dot) denotes the concatenation of sub-labels.

A first option for building those components is to propose the following sets:

- POS={ADJ, ADV, CH, CONJCOO, CONJSUB, DET, INT, INT, MI, N, PREP, PRES, P, PP, V}
- Genre={M, F}; Pers={1, 2, 3}; Num={S, P}
- Mode_Tense={CON, IMP, SUB, IND, INDF, INDI, INDP, INF, PARP, PARPRES}
- Det_Pro={IND, DEM, DEF, POSS, PER, INT}

Each of those components can be learned independently. However, some of them are still mutually exclusive: for example, Person and Gender can be grouped together since their values (respectively in $\{1, 2, 3\}$ and $\{M, F\}$) never occur together. On the contrary, Gender and Number can not be grouped, since, for instance, the value 'F' may occur with 'S' or 'P' within the same label. We end up working with 4 components: $G0$ = POS, $G1$ = Genre $\cup$ Pers $\cup \{\epsilon\}$, $G2$ = Num $\cup \{\epsilon\}$, and $G3$ = Mode_Tense $\cup$ Det_Pro $\cup \{\epsilon\}$. with ϵ the empty string, neutral element for the concatenation. Each of these label subsets can now be learned independently by a specific CRF. In that case, the final label proposed by the system results from the concatenation of all the CRF outcomes. If we denote by $\cdot$ (dot) the concatenation operator, the cartesian product $G0 \cdot G1 \cdot G2 \cdot G3$ generates every L2 label. But it also generates labels which are not linguistically motivated. For example, ADVMP = ADV $\cdot$ M $\cdot$ P $\cdot \epsilon$ is meaningless, since an adverb is invariable. In order to avoid that problem, we have tested two different approaches. The first one consists of using a new CRF, whose features are the components learned independently. The second one consists of intoducing explicit symbolic rules in the concatenation process, in order to rule out the illegal combinations. Example rules are as follows:

- ADV, CONJCOO, CONJSUB and INT can only combine with ϵ
- V can not combine with values of Det_Pro
- DET can not combine with values of Mode_Tense

Those rules exploit the fact that the POS category (*i.e.* the G0 component) is learned with strong enough confidence ($F_1 = 0.94$) to constrain the other sub-labels with which it may combine. We have carried out the tests presented below:

Expe. 1 $CRF(G_i|feat(\text{word}, \text{lemma}, D_3(\text{word}))) \rightarrow ResGi$

We have also tested different versions where feature generation is achieved without lemma but with word tails instead (as in Test IVbis).

Expe. 2 $CRF(G_i|feat(\text{word}, D_3(\text{word}), D_2(\text{word}), D_1(\text{word}))) \rightarrow ResbisGi$

Test VII $CRF(L2|feat(\text{word}, \text{Rlemma}, ResG0, ResG1, ResG2, ResG3)) \rightarrow ResL2$

Test VIIbis *Same as VII, but without lemmas:*
$CRF(L2|feat(\text{word}, ResbisG0, ResbisG1, ResbisG2, ResbisG3)) \rightarrow ResbisL2$

Test VIII *Here, the outcome from Test VII is replaced by symbolic combination rules using the results ResGi.*

Test VIIIbis *Same as VIII, except that the combination rules use ResbisGi.*

Figure 3 illustrates the two possible strategies for recombining the full labels, along with the results from learning each component independently (accuracy measures). Note that the component G2 is better learned without lemmas but

Component	ResGi	ResbisGi
G0	0.94	0.92
G1	0.92	0.92
G2	0.93	0.95
G3	0.95	0.94

Test	VII	VIIbis	VIII	VIIIbis
F_1	0.89	0.875	0.9	0.895

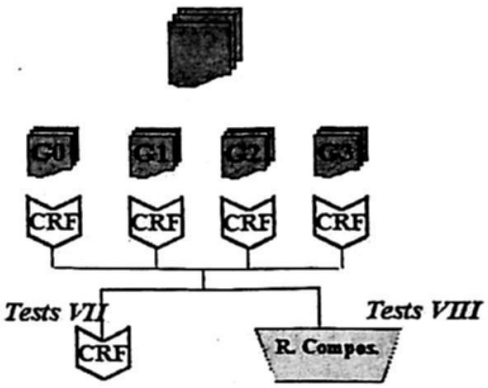

Fig. 3. Learning components: accuracy measures for two recombination strategies.

with word tails. The recombination strategy based on CRF (Test VII and VIIbis) does not improve the scores obtained by direct learning of the full labels on L2 (Test IV and IVbis). However, the rule-based recombination strategy (Test VIII and VIIIbis) does improve direct learning. Test VIIIbis illustrates that, in general, the absence of lemmas can be balanced by word tails associated with symbolic recombination of the labels. Meanwhile, timewise the learning phase is considerably improved by the recombination strategy: Test VIII only takes 75 min., while Test IV takes up to 15h. (using a standard PC). It should also be noted that since the labels obtained by recombination are most of the time only partially (in)correct, those experiments would be better evaluated with other measurements than accuracy.

Note, as well, that the definition we provide of a specific set of labels prevents comparing our performances against those of other taggers.

4 Conclusion

In that paper, we have shown that it is possible to efficiently learn a morpho-syntactic tagger specialised for a specific type of corpus. First, we have seen that the specificities of spoken language are difficult to enumerate. Instead of trying to rule them all, it is natural to rely on Machine Learning techniques. Our experiments all take the input data as they are, without filtering out any difficulties.

Note that it is not possible to rigourously compare the performance achieved by Cordial with the performances reported here, since the target label sets are different. Yet, the performances of the best learned taggers seem comparable to those usually obtained by Cordial on oral corpora. The insentive of using CRF for that task is that it does not require many parameters to be set, and that the settings involved are flexible enough to integrate external linguistic knowledge. We have mostly used here our understanding of the labels in order to focus on learning sub-labels, which are simpler and more coherent. The performance would have also certainly been improved by the use of a dictionary of labels for each word, or each lemma, to specify features.

Finally, it seems quite difficult to still improve the quality of the learned labels by simply relying on the decomposition in simpler sub-labels. However, that strategy by decomposition is very efficient timewise, and the learning process has been greatly improved in that respect. It is also interesting to notice that the most efficient strategy relies on a combination between statistic learning and symbolic rules. Further works are going in that direction.

References

1. Dister, A.: De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelle orale VALIBEL. Thèse de doctorat, Université de Louvain (2007)
2. Valli, A., Veronis, J.: Étiquetage grammatical des corpus de parole : problèmes et perspectives. Revue française de linguistique appliquée **IV**(2) (1999) 113–133
3. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of ICML'01. (2001) 282–289
4. Sutton, C., McCallum, A. In: 1 An Introduction to Conditional Random Fields for Relational Learning. The MIT Press (2007)
5. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of CoNLL. (2003)
6. Pinto, D., McCallum, A., Lee, X., Croft, W.: Table extraction using conditional random fields. In: SIGIR'03: Proceedings of the 26th ACM SIGIR. (2003)
7. Altun, Y., Johnson, M., Hofmann, T.: Investigating loss functions and optimization methods for discriminative learning of label sequences. In: Proceedings of EMNLP. (2003)
8. Sha, F., Pereira, F.: Shallow parsing with conditional random fields. In: Proceedings of HLT-NAACL. (2003) 213–220
9. Jousse, F.: Transformations d'abres XML avec des modèles probabilistes pour l'annotation. Thèse de doctorat, Université de Lille (2007)
10. Zidouni, A., Glotin, H., Quafafou, M.: Recherche d'Entités Nommées dans les Journaux Radiophoniques par Contextes Hiérarchique et Syntaxique. In: Actes de Coria. (2009)
11. Blanche-Benveniste, C., Jeanjean, C.: Le Français parlé. Transcription et édition, Paris (1987)
12. Blanche-Benveniste, C.: Les aspects dynamiques de la composition sémantique de l'oral. In Condamines, A., ed.: Sémantique et corpus. Lavoisier, Hermès, Paris (2005) 39–74