Adapt a Text-Oriented Chunker for Oral Data: How Much Manual Effort is Necessary?

Isabelle Tellier*, Yoann Dupont*, Iris Eshkol** and Ilaine Wang*

* Lattice, University Paris 3 - Sorbonne Nouvelle, ** LLL, University of Orléans

Abstract. In this paper, we try three distinct approaches to chunk transcribed oral data with labeling tools learnt from a corpus of written texts. The purpose is to reach the best possible results with the least possible manual correction or re-learning effort.

1 Introduction

The annotation of transcribed spontaneous speech is a difficult task, because oral corpora are full of irregularities and *disfluencies*. In this paper, we are mainly interested in the task of chunking transcribed oral data. A "chunk" is a non-recursive constituent of linguistic units [Abn91]. The purpose of a chunker is thus to identify the sentence constituents without specifying their internal structure and their syntactic function. This analysis relies on POS labels and can be considered as the best possible that can be reached for such data, for which it is not always possible to provide a full syntactic parsing.

To build a POS tagger and a chunker, several strategies can be considered. The main two options are either writing rules by hand or using supervised machine learning techniques on labeled data. We favor the machine learning approach, which requires less effort and performs better. Unfortunately, for many languages (this is the case for French, the language we are interested in), annotated transcribed oral corpora are rare. We are in a situation where a large corpus of fully annotated written sentences is available, whereas only a small corpus of annotated transcribed oral data is. Our annotated oral data are not large enough to learn a specific POS tagger, but learning a chunker requires less data. Is it worth doing it? The problems we address in this paper are thus the following: to chunk transcribed data, is it better to use a chunker learnt from a large corpus of written texts or one learnt from a small but specific sample of the target transcribed corpus? How much improvement can a limited manual effort bring? To address these questions, we propose three distinct protocols with increasing amount of human intervention, and compare their effectiveness.

The first section of this paper describes the chunking task, its specificities for transcribed oral data and the machine learning technique used. The second section is dedicated to the labeled corpora at our disposal: a large corpus of written texts and a small corpus of transcribed oral data, and their corrected versions. In the last section, we propose three strategies to chunk the transcribed data with various levels of manual adaptation and compare their results.

2 The Chunking Task

2.1 Chunking Transcribed Oral Data

Chunkers, also called shallow parsers, are well adapted for transcribed oral data in which "sentences" are not fully syntactically correct. Some software tools provide this type of analysis but they usually don't behave well on oral data. The reasons are the lack of punctuation marks and the *disfluencies*, which are standardly seen as positions in the speech flow where the linearity is broken. They are very numerous in spoken texts and of various types: repetitions (e.g. "la la" in French, i.e. "the the"); immediate self-corrections (e.g. "le la", i.e. "the (masculine) the (feminine)"); false start (e.g. "les dans" i.e. "the in"), word fragments (e.g. "vous v- vous", i.e."you y- you").

Some attempts to build a chunker specifically adapted to French transcribed data have been developed in France. Most of them consisted in iteratively applying hand-written finite-state transducers, together with lexical and syntactic resources. [BCDW10] tried to automatically annotate French corpora of spontaneous speech transcriptions in super-chunks, i.e. chunks containing complex multiword units. Their parsing was based on a preprocessing stage of the spoken data consisting in reformatting and tagging utterances containing disfluencies. A similar approach has been conducted in [VV99] for POS tagging. [AMF08] proposed another strategy by including a post-correction stage in order to deal with chunking errors due to disfluencies.

Our approach is different. Following [BB05], we believe that disfluence phenomena should be included in the analysis of language even if it raises specific processing issues. To deal with real data and avoid *ad hoc* handmade programs, we favor a machine learning strategy. We detail in the following the kind of learning models we use. A similar strategy has already been applied for POS labeling of French transcribed oral data in [TETP10], but it was based on a different tagset than the one used here.

2.2 Machine Learning for POS labeling and Chunking

To perform both POS labeling and chunking, we used the state of the art machine learning approach for annotation tasks: Conditional Random Fields (or CRFs). As it has often been observed [SP03,CT12] they behave very well on this task.

Introduced in [LMP01], CRFs belong to the family of graphical models. When the graph is linear (which is most often the case), the probability distribution that the annotation sequence y is associated with the input sequence x is:

$$p(y|x) = \frac{1}{Z(x)} \prod_t \exp\Big(\sum_{k=1}^{K} \lambda_k f_k(t, y_t, y_{t-1}, x)\Big)$$

Where $Z(x)$ is a normalization factor. This computation is based on K features f_k (usually binary functions), provided by the user. The feature f_k is activated (i.e. $f_k(t, y_t, y_{t-1}, x) = 1$) if a configuration occurring at the current position t

in the sequence, concerning y_t, y_{t-1} (i. e. the values of the annotation at the positions t and $t-1$) and x is observed. Each feature f_k is associated with a weight λ_k, estimated during the learning step. The most efficient implementation of linear CRFs is Wapiti[1], which uses L1 and L2 penalizations to select the best features during the learning step [LCY10]. It is the software we have used.

3 Corpora and Labeling Conventions

In this section, we first describe two French corpora at our disposal. The first one, called FTB (French TreeBank), is a treebank of written sentences, which can thus be easily transformed into a labeled corpus for POS and chunking. The second one, called ESLO 1 (Enquête Sociolinguistique d'Orléans[2]), is made of transcribed oral data. Originally, it was not at all labeled.

Our basic methodology consists in applying labeling tools learnt from the corpus of written sentences to the corpus of transcribed oral data. To measure the effectiveness of these tools, we had to build a reference labeled version of the transcribed corpus manually. Some linguistic choices had thus to be made to take into account the specificities of our transcribed data while remaining as much as possible compatible with the labels of the written sentences. To do so, we defined some labeling conventions, especially concerning the disfluencies. In fact, we produced two reference labeled versions of the ESLO 1 corpus tagged with the labels of the FTB, with an increasing level of adaptation. We describe these two variants (corpus 1 and corpus 2) in the last two subsections.

3.1 French TreeBank (FTB)

The variant of the FTB we used is made of about 10 000 fully parsed sentences extracted from articles of the newspaper "Le Monde" [ACT03]. The set of 30 POS tags is described in [CC08]. The distinct possible kinds of chunks, together with the possible POS tags corresponding to their head, are the followings: AP for adjectival chunk (ADJ, ADJWH), AdP for adverbial chunk (ADV, ADVWH, I for interjection), CONJ for conjunction chunk (CC, CS), NP for nominal chunk (CLO, CLR, CLS, NC, NPP, PRO, PROREL, PROWH), PP for prepositional chunk (P, P+D, P+PRO), VN for verbal chunk (V, VIMP, VINF, VPP, VPR, VS). To transform a chunk analysis into a word annotation, we use the classical BIO labeling format (B for Beginning, I for In, O for Out is useless here since every word is member of a chunk).

3.2 ESLO 1

The ESLO 1 campaign gathered a large oral corpus among which we extracted for our experiments a sub-corpus of 8093 graphical words (i.e. tokens between

[1] http://wapiti.limsi.fr/
[2] Sociolinguistic Survey of Orléans

two separators) belonging to 852 speaking turns. The main principles of our transcription guidelines are those followed by the researchers of the domain. First, words are transcribed using their standard spelling. Transcriptions do not contain any punctuation marks because the notion of sentence is not considered relevant [BBJ87]. The texts of ESLO 1 display disfluencies phenomena that are specific to spoken language: examples of repetitions, self-corrections, truncations, etc. it contains are provided in the following.

3.3 Labeling ESLO 1 with the FTB labels: corpus 1

We describe here how we choose to label oral disfluencies.

Repetition: When a token is repeated, the POS tags of both tokens are the same but, at the chunk level, two cases are possible:
• if the repeated token is a chunk head, then two distinct chunks are defined:
(et/CC)$_{CONJ}$ (et/CC)$_{CONJ}$ (elle/CLS)$_{NP}$ (me/CLO)$_{NP}$ (disait/V)$_{VN}$[3]
• in every other case, both tokens belong to the same chunk:
(la/DET la/DET jeune/ADJ fille/NC)$_{NP}$[4]

Discourse markers are considered as interjections (I) and put into an adverbial chunk: (on/CLS)$_{NP}$ (peut/V)$_{VN}$ (commencer/VINF)$_{VN}$ (bon/I)$_{AdP}$ (alors/I)$_{AdP}$[5]

False starts and word fragments which are impossible to interpret are labelled as interjections and are part of an adverbial chunk: (c'/CLS)$_{NP}$ (est/V)$_{VN}$ (difficile/ADJ)$_{AP}$ (heu/I)$_{AdP}$ (les/I)$_{AdP}$ (dans/P ma/DET classe/NC)$_{PP}$[6]

Others interruptions of the morpheme being enunciated are interpreted according to the context:
• (vous/PRO)$_{NP}$ (êtes/V)$_{VN}$ (in-/NC)$_{NP}$ (institutrice/NC)$_{NP}$[7]
• (chez/P vous/PRO)$_{PP}$ (chez/P v-/PRO)$_{PP}$[8]

3.4 An Adapted Version of the Labeling for Oral Data: corpus 2

To go further, it is also possible to define new specific chunks to treat disfluencies. To build the corresponding reference corpus, we made some new choices.

UNKNOWN POS tag and chunk: The UNKNOWN tag exists in the FTB, where it is assigned to foreign words. In corpus 2, we choose to also assign it to false starts, word fragments and orthographic errors of transcriptors. It is thus

[3] and and she told me
[4] the the young girl
[5] we can start well then
[6] it is difficult er the in my classroom
[7] you are schoo- schoolteacher
[8] at your's at y-

both a POS tag and a new kind of chunk.

Interjection chunk (IntP) is also added as a new kind of chunks, used for interjection phrases and discourse markers: (c'/CLS)$_{NP}$ (est/V)$_{VN}$ (difficile/ADJ)$_{AP}$ (euh/I)$_{IntP}$ (les/UNKNOWN)$_{UNKNOWN}$ (dans/P ma/DET classe/NC)$_{PP}$[9]
Interjections and discourse markers can be part of a nominal chunk in the case of hesitations inside a NP like in:
• (l'/DET école/NC euh/I publique/ADJ)$_{NP}$[10]
• (des/DET hm/I inconvénients/NC)$_{NP}$[11]
unlike those *following* the nominal chunks: (des/DET idées/NC laïques/ADJ)$_{NP}$ (quoi/I)$_{IntP}$[12]

4 Three Experiments

We now describe in details three distinct experiments we have conducted to label our transcribed oral data and the results obtained. These experiments require an increasing amount of manual effort. In the following, the evaluation of the chunks is done using the strict equality criterion, meaning that two chunks are considered equal if and only if they share exactly the same frontiers and type.

4.1 First Approach: Direct Use of Written Texts Oriented Tools

The first and most simple strategy consists in applying the POS tagger and the chunker learnt from the FTB in cascade, without any adaptation, on the transcribed oral data of ESLO 1 (with corpus 1 as reference labelled data).

Figure 1 shows the templates used to define the features of the CRF for the POS tagging and for the chunking. For the POS tagging (on the left), we also used an external resource called the LeFFF (Lexique des Formes Fléchies du Français[13]) which is integrated into the CRF as a set of boolean attributes, one per distinct POS tag, representing whether or not the word is associated to this specific tag in the LeFFF. For the chunker (on the right) the features of the CRF are based on the correct POS tags of the FTB.

Figure 2 shows the results of these first experiments. The evaluations on the FTB are made by a 10-fold cross-validation. We evaluate the chunking with the micro- (resp. macro-) average of the F-measures of the obtained chunks, which is the average of the F-measures of every kind of chunk weighted (resp. not weighted) by their frequencies. Note that for corpus 1, the chunker is applied in cascade after the POS tagger, whereas it is based on the correct POS tags for the FTB. As expected, for oral data the quality loss is important on POS tagging (nearly 17%) and even worse when cascading the chunker afterwards. These poor results justify manual effort to adapt the tools.

[9] it is difficult er the in my classe
[10] the public er school
[11] some hm disadvantages
[12] laicist ideas what
[13] French Inflected Forms Lexicon

Feature	Type on y	window on x
Starts with upper ?	Unigram	[-2 .. 1]
Is punctuation ?	Unigram	[-2 .. 1]
Is a decimal ?	Unigram	[-2 .. 1]
3 last word's letters	Unigram	[-2 .. 1]
LeFFF information	Unigram	[-2 .. 1]
positional annotation	Bigram	$\emptyset$

Feature	Type on y	Window on x
Word	Unigram	[-2 .. 1]
POS	Bigram	[-2 .. 1]

Fig. 1. Feature Templates for POS and Chunk Learning resp. on the FTB

corpus	accuracy of the POS	micro-average of the chunker	macro-average of the chunker
FTB	97,33%	97,53	90,4
corpus 1	80,98%	77,24	76

Fig. 2. Results of the First Approach

4.2 Second Approach: Manual Correction of the POS Labeling

The second approach integrates a manual intervention after the application of the POS tagger, to correct the tags assigned to the transcribed corpus. The chunker learnt from the FTB is then applied on this corrected version of the oral data. This process is displayed on Figure 3, the manual effort being in bold. This correction is meant to compensate errors made by the POS tagger which typically depend on differences between written and oral productions. For example, "bon"[14] is used 99% as an adjective in the FTB, whereas it is much more frequently (83%) an interjection in corpus 1, "oui" and "non" [15] are labeled as ADV, I or NC in the FTB, while they are only interjections in corpus 1 (the reference corpus for this experiment). In this approach, the chunker is applied on correct POS tags for the oral data, as previously on the FTB. 1593 POS tags out of 8093 had to be manually corrected. The new micro-average of chunks on corpus 1 is then 87,74 while the new macro-average is 88,43. We have a significant improvement of results.

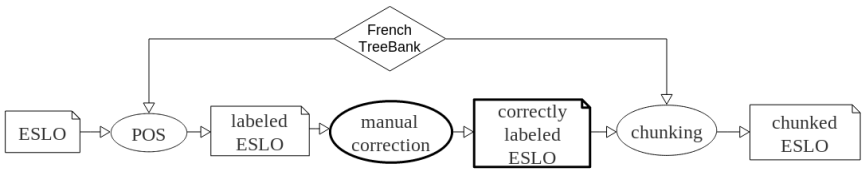

Fig. 3. Protocol of the Second Approach

[14] "well"
[15] "yes" and "no"

4.3 Third Approach: Learning of a Specific Chunker for Oral Texts

The last approach consists in learning a new chunker from the corrected transcribed speech data. We have already noted that learning a POS tagger would require a larger amount of data (i.e. a stronger manual labeling effort). The situation is different for a chunker, which mainly relies on the POS tags. The POS tags are much less numerous than words, and are thus much more likely to provide useful repetitions. Furthermore, this strategy gives the opportunity to define oral-specific chunks, as explained in section 3.4. The process used is displayed on Figure 4. The reference corpus for this experiment is thus corpus 2. The manual correction concerns 902 chunk tags out of 8093.

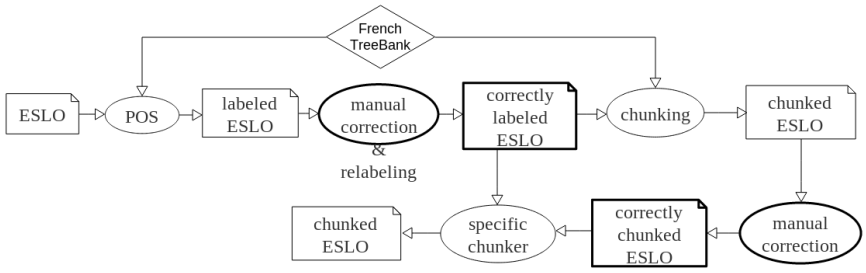

Fig. 4. Protocol of the third experiment

Table 5 shows the information used to train the CRF for the oral-specific chunk labeling task and the results obtained by 10-fold cross-validation.

Feature	Type on y	Window on x
Word	Unigram	[-2 .. 0]
POS	Unigram & Bigram	[-2 .. 1]
POS couple	Unigram	{-2,0} & {-1,0}

F-measures	on corpus 2
micro	96.65
macro	96.08

Fig. 5. Feature templates and F-measures on fully corrected ESLO

5 Observations and Conclusion

Analyzing the results, some observations can be made. First, the major improvements between the second and the third experiments concern adverbial (AdP) and nominal (NP) chunks. This is due to the high number of AdPs in the oral corpus, because of the interjections it contains. The situation is different for the conjunctive chunk (CONJ), equally well treated by both experiments. Finally, Adjectival (AP) and verbal chunks (VN) are better treated by the second experiment, probably because they are more frequent in written data.

Our oral corpus is characterized by a very significant number of interjections and discursive markers. The introduction of the new IntP chunk in corpus 2 reduces the number of adverbial chunks comparatively to corpus 1 and induces a significant improvement of F-measure for this chunk: AdP has a F-measure of 58,14 in the first experiment, 71,87 in the second one and 95,76 in the third one! When the POS I is correct, the IntP is very well identified: its F-measure is 99,4.

The remaining errors concern either repetitions which are not correctly treated or interjections inside nominal chunks, like the example of section 3.4 wrongly chunked as: (l'/DET école/NC)$_{NP}$ (euh/I)$_{I}$ (publique/ADJ)$_{AP}$ (see note 10).

More generally, we see that an excellent POS tagger learnt from written sentences makes about 20% errors on transcribed data, and induces the same level of errors for the chunker. Correcting the POS errors helps the chunker by improving its score of about 10%. But learning a specific chunker from corrections at both levels is even better, even with limited training data.

References

[Abn91] S Abney. Parsing by chunks. In R. Berwick, R. Abney, and C. Tenny, editors, *Principle-based Parsing*. Kluwer Academic Publisher, 1991.

[ACT03] A. Abeillé, L. Clément, and F. Toussenel. Building a treebank for french. In A. Abeillé, editor, *Treebanks*. Kluwer, Dordrecht, 2003.

[AMF08] J-Y. Antoine, A. Mokrane, and N. Friburger. Automatic rich annotation of large corpus of conversational transcribed speech: the chunking task of the epac project. In *Proceedings of LREC'2008*, may 2008.

[BB05] C. Blanche-Benveniste. *Sémantique de l'oral*, chapter Sémantique et corpus. Les aspects dynamiques de la composition sémantique de l'oral. 2005.

[BBJ87] C. Blanche-Benveniste and C. Jeanjean. *Le français parlé, transcription et édition*. Didier Erudition, 1987.

[BCDW10] O. Blanc, M. Constant, A. Dister, and P. Watrin. Partial parsing of spontaneous spoken french. In *Proceedings of LREC'2010*, 2010.

[CC08] B. Crabbé and M. H. Candito. Expériences d'analyse syntaxique statistique du français. In *Actes de TALN'08*, 2008.

[CT12] M. Constant and I. Tellier. Evaluating the impact of external lexical resources unto a crf-based multiword segmenter and part-of-speech tagger. In *Proceedings of LREC 2012*, 2012.

[LCY10] Thomas Lavergne, Olivier Cappé, and François Yvon. Practical very large scale CRFs. In *Proceedings of ACL'2010*, pages 504–513. Association for Computational Linguistics, July 2010.

[LMP01] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML 2001*, pages 282–289, 2001.

[SP03] Fei Sha and Fernando Pereira. Shallow parsing with conditional random fields. In *Proceedings of HLT-NAACL*, pages 213–220, 2003.

[TETP10] I. Tellier, I. Eshkol, S. Taalab, and J. P. Prost. Pos-tagging for oral texts with crf and category decomposition. *Research in Computing Science*, 46:79–90, 2010.

[VV99] A. Valli and J. Veronis. Etiquetage grammatical des corpus de parole: problèmes et perspectives. *Revue française de linguistique appliquée*, 4(2):113–133, 1999.