Chunk Different Kind of Spoken Discourse: Challenges for Machine Learning

Iris Eshkol-Taravella, Mariame Maarouf, Flora Badin, Marie Skrovec, Isabelle Tellier
MoDyCo UMR7114, LLL UMR7270, Lattice UMR8094
200 avenue de la République, 92001 Nanterre, France,
10 Rue de Tours, 45065 Orleans, France
1 rue Maurice Arnoux, 92120 Montrouge, France,
ieshkolt@parisnanterre.fr, maarouf.mariame@gmail.com, {marie.skrovec, flora.badin @univ-orleans.fr}

Abstract

This paper describes the development of a chunker for spoken data by supervised machine learning using the CRFs, based on a small reference corpus composed of two kinds of discourse: prepared monologue vs. spontaneous talk in interaction. The methodology considers the specific character of the spoken data. The machine learning uses the results of several available taggers, without correcting the results manually. Experiments show that the discourse type (monologue vs. free talk), the speech nature (spontaneous vs. prepared) and the corpus size can influence the results of the machine learning process and must be considered while interpreting the results.

Keywords: chunking, machine learning, CRF, automatic segmentation of spoken data, oral corpus, kind of spoken discourse

1 Introduction

The notion of sentence is generally considered as irrelevant to the analysis and treatment of spoken data (Blanche-Benveniste et al, 1990; Fribourg Group, 2012). Researchers suggested different segmentation units for spoken data as part of projects such as Rhapsody or Orfeo. The project SegCor[1] aims to segment the data of talk-in-interaction at different levels. Its first level segmentation concerns minimal syntactic units, which are called chunks.

A "chunk" is a non-recursive constituent of linguistic units (Abney 1991). Chunking (or shallow parsing) identifies the surface syntactic structure of a sentence and it can be done automatically. The purpose of a chunking is to identify constituents of the sentence without specifying their internal structure and their syntactic function, which is based on previous morphosyntactic labeling. Chunking is a good way to perform syntactic analysis automatically on spoken data, for which it is not always feasible to provide a full syntactic parsing. Chunkers are well adapted for transcribed oral data in which “sentences” are not always syntactically fulfilled. The chunk is supposed to be a relevant unit for spontaneous speech. Blanche-Benveniste (1997) has shown that it is in the chunks where the reparation markers often occur in spoken data. Some software tools provide this type of analysis but their performance is usually low on oral data.

There are several strategies to develop a chunker. There have been attempts to build a chunker that is particularly adapted to French data by using symbolic methods (Blanc et al, 2008, 2010, Antoine et al, 2008). The method consists of iteratively applying finite-state transducers, together with lexical and syntactic resources. The supervised machine learning seems to be particularly effective in this task as shown by (Sha and Pereira, 2003 Tellier et al, 2012, 2014, Tsuruoka et al, 2009). The present research continues the work of (Tellier et al, 2014) and uses the method of supervised learning. Oral data are characterized by discursive variety: situational variety (private conversation, public debate ...), language tasks (explain, narrate, describe ...), genres (travel stories, interviews ...) or register (common, familiar…). The nature of the data influences and guides the learning process. In (Tellier et al, 2014), the labeled reference corpus consisted of sociolinguistic interviews; in this work, we were interested in two other communicative situations: a university conference and a spontaneous discussion between friends during a dinner. Our objective is to develop a chunker for spoken data by using Conditional Random Fields (CRFs). We want to find out how the discourse type (prepared monologue vs. spontaneous talk in interaction) can influence the results of the experiments and which features are most relevant to each communicative situation.

2 Reference corpus constitution

We dispose of two corpora for spoken French: ESLO2[2] and CLAPI[3]. We selected two types of speech: a conference, i.e. a prepared monologue (10 minutes, 2120 tokens) from the corpus ESLO2 (M) and a discussion between three people, a spontaneous interaction taking place in a private context (10 minutes, 2461 tokens) from the corpus CLAPI (D).

2.1 Pretreatment

The two files used in this work were segmented and annotated by Treetagger (Schmid, 1994) and Dismo (Christodloulides et al., 2014). Multiword units were

[1] A Franco-German project, funded by the National Research Agency (ANR-15-FRAL-0004)

[2] ESLO : Enquêtes Sociolinguistiques à Orléans, Sociolinguistic Survey of Orléans, http://eslo.huma-num.fr/

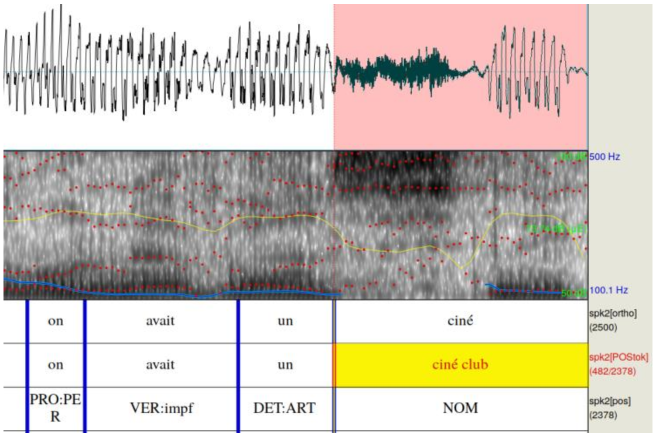
[3] Corpus of spoken language in interaction, http://clapi.ish-lyon.cnrs.fr/

identified by Lefff (Sagot et al., 2010). The result of the preprocessing is shown in Figure 1.

2.2 Chunks typology

Our typology, based on a previous typology presented in Tellier et al. (2014), was complemented by two new labels (FNO and ARTIC). It contains nine categories:

– adjectival chunk (AP): adjective head after the verb (*it is too pretty*);
– adverbial chunk (AdP): syntagma whose head is an adverb (*perhaps*);
– nominal chunk (NP): noun phrases including adjectives placed before and after the name and non-clitic pronouns (*your beautiful shoes*);
– prepositional chunk (PP): syntagma introduced by a preposition (*by far*);
– verbal chunk (VP): phrases organized around a verbal head, associated with its clitics (*we hear you – nous vous entendons*);
– punctuation (SENT): typographical marks ;
– articulator (ARTIC): category which includes all kind of cohesive linking and organizing markers on different structural levels of spoken data as relative pronouns, conjunctions, discourse markers, etc. (*and*, *that*, *which*, *but*);
– nucleus forms (FNO): inspired by the work of Benzitoun et al. (2012), this category includes autonomous elements constituting illocutionary units (*yes*, *no*, *shit*, *hello*);
unknown (UNKNOWN): a category for unidentified chunks like false starts, misspelled words, etc.



on	avait	un	ciné	spk2[ortho] (2500)
on	avait	un	ciné club	spk2[POStok] (482/2378)
PRO:PER	VER:impf	DET:ART	NOM	spk2[pos] (2378)

Figure 1 : Result visualized by Praat of the segment "at primary school we had a film club"[4]

2.3 Manual annotation

Two researchers annotated the pretreated corpus according to the established typology. The manual annotation is evaluated by the inter-annotator agreement. The inter-annotator agreement was calculated using the Kappa Cohen measurement (Cohen, 1960) and got a very good score (88) according to (Landis and Koch, 1977*)*. We provided a compromise version for the corpus. This third annotation was used as reference corpus for the machine learning evaluation. The annotation was realized using the Praat software (Boersma and Van Heuven, 2001) and the BILOU format[5] (Ratinov and Roth, 2009). This format delimits not only as a single unit but also indicates the position of a word within its unit. By using Praat, the annotators can listen to the recordings and therefore understand the situations better. The annotated corpus contains 1069 chunks in M and 1455 chunks in D. In the two corpora, the annotated chunks are unevenly distributed (PP forming a large proportion of 30% in M vs. 11% in D, while VP representing 40% in D vs. 23% M etc.).

3 Machine learning

The machine learning process aims to indicate the borders of each chunk and to determine its type. The reference corpus has a small size, hence we choose the CRF models (Conditional Random Fields) (Lafferty et al., 2001) that has already shown a good performance for this task (Sha and Pereira, 2003 Tellier et al, 2012, 2014, Tsuruoka et al, 2009). We applied the chunking to the POS labeled corpus. Tellier et al. (2014) showed that it is possible to train a chunker for spoken corpus with non-corrected POS tags and with a small size of reference corpus. The authors obtained a micro-average of 88%. We continue the same approach but with a methodology that we redefined according to the specificities of oral data : (1) our data is more heterogeneous because it includes two types of oral discourse; (2) human annotators use audio sound to determine annotation choices; (3) the set of labels was modified (two new labels ARTIC and FNO were added); (4) the results of morphosyntactic labels suggested by several taggers are added as features and integrated into the CRF model.

We tested four taggers: TreeTagger (Schmidt, 1994); SEM (Tellier et al, 2012.) exploited by (Tellier et al, 2014) and using morphosyntactic labels of (Crabbe et al., 2008); syntactic dependencies parser (Kahane et al, 2017) developed by researchers of Orfeo project ; Perceo (Benzitoun et al., 2012), POS tagger for oral data using FNO label which is also present in our chunks typology.

We performed the experiments on three corpora: ESLO2 (M), CLAPI (D), ESLO2 and CLAPI (M + D). The aim is to check if speech type (monologue / discussion between three people), speech nature (spontaneous / prepared) and corpus size can influence the machine learning results. We tested many configurations by combining and varying templates

[4] Praat is a transcription and manual annotation tool for oral data (http://www.fon.hum.uva.nl/paul/praat.html).

[5] B, beginning, first token of chunk; I, inside, an element within a chunk; L, last, a last element of chunk; O, out, an element outside the chunk; U, unit, a chunk composed of one token.

[token + POS][6]. The features of the CRF were based on the POS of three tools TreeTagger, Perceo and Orfeo. For Orfeo parser, we tried two additional combinations (1) POS label for current token, (2) POS label for current token and its head. First, for each combination, the current line token was tested. Then, for each corpus, the three combinations giving the best were selected for the test. We also included the same combinations on token+1 and token-1. After selecting the best result, other columns were added such as lemma, which resulted from TreeTagger tagging. The Figure 2 shows the patterns of the best combinations for each corpus.

4 Results and evaluation

The evaluations on three corpora (M, R, M + R) were made by a 10-fold cross-validation. We evaluate the chunking with the micro-average of the F-measures of the obtained chunks, which is the average of the F-measures of every type of chunk weighted by their frequencies.

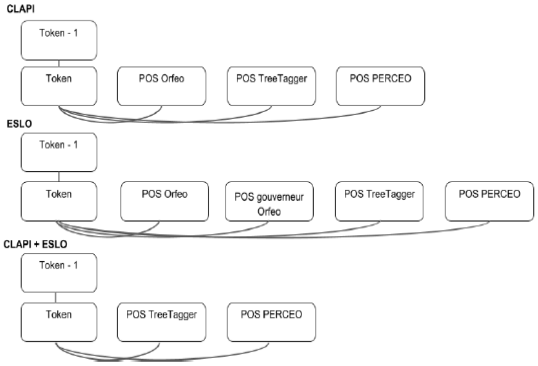


Figure 2 : Best feature templates for each corpus

For the corpus D, the best results are templates composed of TreeTagger, PERCEO, Orfeo POS and current and preceding tokens (83.2%). The best score for the corpus M was obtained by a similar combination but this time taking into account the POS of the head proposed by syntactic dependencies parser of Orfeo (85.8%).

The monologue, part of the conference, contains longer utterances and no interaction, this is the reason why dependency links are more present in this corpus.

In the case of the corpus M + D, the best results were obtained by using only the POS of TreeTagger and Perceo with current and preceding tokens (85.7%). These results show that the use of labels proposed by two tools learned on oral data: PERCEO and Orfeo parser, in templates is relevant. The corpus size is also an important criterion because we can see in our results that longer corpora don’t need a lot of POS labels and the combination of the results of two tools is enough. It is surprising to note that TreeTagger seems to be more relevant in this case than the Orfeo parser, which is developed for spoken data. The corpus M, a prepared monologue, gives the best machine learning results than the corpus D, the spontaneous discussion.

	M	D	M + D
micro-average	85.8%	83.2%	85.7%
POS TreeTagger	x	x	x
POS Perceo	x	x	x
POS Orfeo	x	x	
Governor Orfeo	x		
Token Token-1	x	x	x

Figure 3 : Better results of micro-average

The evaluation of labels proposed by our chunker shows that the FNO label gets worse results (23.52% F-measure). Indeed, some tokens are ambiguous, like *yes (oui, ouais), no* which can be *FNO* or *ARTIC* (discourse markers). Thus, *yes (ouais)* in response to a question will be considered as autonomous predicate (a sentence like word) and therefore annotated (FNO), like here:

Eli je [VP B] vous [VP I] sers [VP L]? (*I serve you)*
BEA ouais [FNO U] (*yes*)

On the other hand, the same form can be in the periphery of the predicate. This form will be considered as non-autonomous discursive articulator, as in the example below, where *yes (ouais)* closes ELI’s speaking turn:

ELI non [ARTIC U] mais [ARTIC U] tu [VP B] sais [VP L] (*no but you know*)
tu [VP B] en [VP I] mets [VP L] pas [AdP B] beaucoup [AdP L] (*you don’t put a lot*)
tu [VP B] en [VP I] mets [VP L] un [NP B] fond [NP L] ouais [ARTIC U] (*you put a depth*)

There are a few other common mistakes. Many chunks NP are annotated as PP because of the ambiguity between the preposition *de* followed by a definite article and the partitive article (*du, de la, etc.*), both have the same form. A quarter of the AP are considered as VP because AP follows often VP. The chunks boundaries (labels B, L, U) are generally better annotated (Figure 4).

[6] token+SEM, token+SEM+TTG, token+SEM+TTG+Orfeo, token+SEM+TTG+Orfeo+Perceo, token+Orfeo, token+Orfeo+TTG, token+1 et token-1.

	B	I	L	U
D	0.94	0.86	0.91	0.94
M	0.92	0.87	0.93	0.9
M + D	0.93	0.86	0.92	0.93

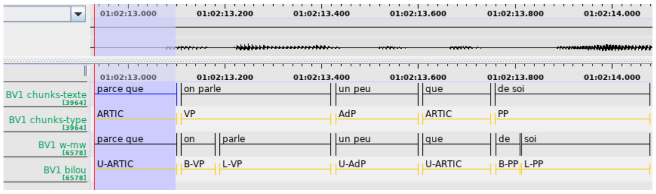
Figure 4 : Results of F-measure for BILU labels

5 Conclusion

The article described the development of a chunker for two different kinds of spoken discourse: a monologue in a conference and a spontaneous discussion between three people during dinner. We trained CRFs by using a small reference corpus. We obtained the best results on the corpus D (83.2%), with templates composed of TreeTagger, PERCEO, Orfeo POS and current and preceding tokens. We obtained the best score for the corpus M (85.8%) with a similar combination, but we also added the governor label proposed by the Orfeo syntactic dependency parser. The experiments showed that the discourse type should be considered while training and interpreting the results. Thus, the results of dependency parsing are more relevant to integrate in CRF model for the monologue in which the utterances are long and dependencies relations are more present. FNO label obtained a better score in a discussion because it is very present in this corpus. In the case of the extended corpus M + D, we obtained the best results (85.7%) by using few features: the POS of only two tools: TreeTagger and Perceo with current and preceding tokens.

The chunker that we developed is launched with the Python language including an interface[7]. It identifies chunks not only from the transcriptions of spoken data without sound but also from the tokens of the transcriptions which are aligned with the sound, by using Jtrans (Cerisara et al., 2009). We obtain the results in Elan software (Brugman and Russel, 2004) format, which shows the transcription and annotation by tiers (Figure 5). There are two ways to visualize the results: 1). the label of the chunk type (eg. ARTIC); 2). the label of the chunk type and the BILOU (eg. U-ARTIC), which are presented side by side. Elan format is convertible to Exmaralda, (Schmidt and Wörner, 2009), a software that is used in the SegCor project. The chunker can be successfully applied also on the CoNLL data sets. It is possible to use our chunker for other tasks. For example, we used the results of chunker for automatic segmentation of transcriptions aligned to the sound in macro-syntactic periods (Kalashnikova et al. 2020).

We have several directions for future work : (1) to add some information from the records as prosody; (2) during manual annotation in cases where human annotators hesitate between different possible labels, to leave both options that will improve the chunker results; (3) to add annotation rules for some recurrent and systematic phenomena, for example that a speaking turn begins with a B or U border; (4) to include in the training corpus the maximum communication's situations to generalize the chunker's development for speech data. Future work may also concern the automatic detection of intonative units and their relations to the chunks. Whereas the chunk is a non-recursive, *microsyntactic* unit organized around a head (nominal, verbal, adjectival, prepositional, etc.), the intonative unit is defined by *prosodic* criteria, i.e. rise or fall of the fundamental frequency, accent prominence, pause, etc. Syntactic and intonative units do not always coincide. In a future work, we could investigate whether the final breath group boundaries coincide with the end of chunks.



BV1 chunks-texte [3964]: parce que | on parle | un peu | que | de soi

BV1 chunks-type [3964]: ARTIC | VP | AdP | ARTIC | PP

BV1 w-mw [6578]: parce que | on | parle | un peu | que | de | soi

BV1 bilou [6578]: U-ARTIC | B-VP | L-VP | U-AdP | U-ARTIC | B-PP | L-PP

Figure 5 : Result from the chunker visualized by Elan

6 Acknowledgements

This work was supported by the ANR project SegCor (ANR-15-FRAL-0004).

This work was made during the internship of Mariame Maarouf, co-supervised by Isabelle Tellier, who left us on June 1, 2018. We would like to pay here our respects and to express our gratitude for his enthusiasm, ideas and advices, without which this article would not have been possible.

7 Bibliographical References

Abney, S. (1991). Parsing by chunks. In R. Berwick, Abney R., & C. Tenny (Eds.), *Principle-Based Parsing*, Kluwer Academic Publisher.

Antoine, JY, Mokrane, A., and Friburger, N. (2008). Automatic Rich Annotation of Large Corpus of Conversational transcribed speech: the Chunking Task of the EPAC Project. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'8), pp. 678–683, Marrakech, Morocco, may. European Language Resource Association (ELRA).

Benzitoun, C., Fort, K., and Sagot, B. (2012). TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. In J. Mauricio Molina Mejia, D. Schwab & G. Sérasset (éds.), Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 3, RECITAL, pp. 99-112.

Blanc, O., Constant, M., Dister, A., and Watrin, P. (2008). Corpus oraux et chunking. In N.Audibert & S.Rosset (éds.), Proceedings of the Joint Conference JEP-TALN 2008, online [http://www.afcp-parole.org/doc/Archives_JEP/2008_XXVIIe_JEP_Avignon/PDF/avignon2008_pdf/JEP/092_jep_1614.pdf].

[7] The tool will be available in Linux at http://segcor.cnrs.fr/deliverable/tools/.

Blanc O., Constant M., Dister A., Watrin P. (2010). Partial parsing of spontaneous spoken french. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), pp. 2111–2116, Valletta, Malta, may. European Language Resource Association (ELRA).

Blanche-Benveniste, C., Bilger, M., Rouget, C., and Van Den Eynde, K. (1990). *Le français parlé*. Études grammaticales, Paris, CNRS Éditions.

Blanche-Benveniste, C. (1997). *Approches de la langue parlée en français*. Paris, Ophrys.

Boersma, P., Van Heuven, V. (2001). Speak and unSpeak with Praat. *Glot International*, 5(9/10):341-347.

Brugman, H., Russel, A. (2004). Annotating multimedia/multi-modal resources with ELAN. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), pp. 2065-2068.

Cerisara, C., Mella, O., and Fohr, D. (2009). Jtrans: an open-source software for semi-automatic text-to-speech alignment. Proceedings of the Tenth Annual Conference of the International Speech Communication Association.

Christodoulides, G., Avanzi, M., Goldman, J-P. (2014). DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 3902–3907, Reykjavik, Iceland, may. European Language Resource Association (ELRA).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1): 37–46.

Groupe de Fribourg, (2012), *Grammaire de la période*, Berne, Peter Lang.

Kahane, S., Deulofeu, J., Gerdes, K., Nasr, A., and Valli, A. (2017). Annotation micro et macrosyntaxique manuelles et automatique de français parlé, *Journée Floral*, mars 2017, Orléans.

Kalashnikova, N., Grobol, L., Eshkol-Taravella, I., Delafonatine, F. (2020). In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020).

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In C.E. Brodley & A.P. Danyluc (eds.), Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), pp. 282-289.

Ratinov, L., Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In S.Stevenson & X.Carreras (eds.), Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009), pp. 147-155.

Sagot B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), pp. 2744–2751, Valletta, Malta, may. European Language Resource Association (ELRA).

Schmidt H. (1994). Probabilistic part-of-speech tagging using decisions trees. Proceedings of International Conference on New Methods in Language Processing, volume 12, pp. 44-49.

Schmidt, T. and Wörner, K. (2009). Exmaralda–creating, analysing and sharing spoken language corpora for pragmatic research. Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA), 19(4), pp. 565–582.

Sha F., Pereira F. (2003). Shallow parsing with conditional random fields. In Eduard Hovy (Conference Chair), et al., editors, Proceedings of the Joint Conference *HLT-NAACL 2003*, pp. 213-220.

Tellier, I., Duchier, D., Eshkol, I., Courmet, A., and Martinet M. (2012). Apprentissage automatique d'un chunker pour le français, In J. Mauricio Molina Mejia, D. Schwab & G. Sérasset (eds.), Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, pp. 421-438.

Tellier, I., Eshkol, I., Dupont, Y., Wang, I. (2014). In P. Blache, F. Béchet, B. Bigi (eds.), Peut-on bien chunker avec de mauvaises étiquettes pos ? Proceedings of *TALN2014*, pp. 125-136

Tsuruoka, Y., Tsujii, J., Ananiadou, S. (2009). Fast full parsing by linear-chain conditional random fields, In A. Lascarides, C. Gardent, & J. Nivre (eds.), Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pp. 790-798.

8 Language Resource References

CLAPI : Corpus of spoken language in interaction, http://clapi.ish-lyon.cnrs.fr/

ESLO : Enquêtes Sociolinguistiques à Orléans, Sociolinguistic Survey of Orléans, http://eslo.huma-num.fr/