Macrosyntactic Segmenters of a French spoken corpus

Ilaine Wang[1], **Sylvain Kahane**[1], **Isabelle Tellier**[2]
[1]MoDyCo, Université Paris Ouest & CNRS, Nanterre, France
[2]LaTTiCe, Université Paris 3 & CNRS, Paris, France
i.wang@u-paris10.fr, sylvain@kahane.fr, isabelle.tellier@univ-paris3.fr

Abstract

The aim of this paper is to describe an automated process to segment spoken French transcribed data into macrosyntactic units. While sentences are delimited by punctuation marks for written data, there is no obvious hint nor limit to major units for speech. As a reference, we used the manual annotation of macrosyntactic units based on illocutionary as well as syntactic criteria and developed for the Rhapsodie corpus, a 33.000 words prosodic and syntactic treebank. Our segmenters were built using machine learning methods as supervised classifiers : segmentation is about identifying the boundaries of units, which amounts to classifying each interword space.
We trained six different models on Rhapsodie using different sets of features, including prosodic and morphosyntactic cues, on the assumption that their combination would be relevant for the task. Both types of cues could be resulting either from manual annotation/correction or from fully automated processes, which comparison might help determine the cost of manual effort, especially for the 3M words of spoken French of the Orfeo project those experiments are contributing to.

Keywords: sentence boundary segmentation, spoken French, machine learning

1. Introduction

The tools we present are developed as part of the Orfeo project, funded by the French National Research Agency (ANR), which aim is to propose syntactic annotations in order to compare the syntax of written and spoken French. We gathered a total of 3 million words of transcription of spoken French.
It is well known that the notion of sentence is not easily applicable to spontaneous speech (Pietrandrea et al., 2014 to appear) and the transcriptions we have are not segmented, which is a real problem for parsing. Based on the results of the Rhapsodie project (Lacheret et al., 2014), we propose a well formalised segmentation for spoken French called macrosyntactic segmentation and we have at our disposal the 33,000-word Rhapsodie treebank. Our approach consists in using this resource and machine learning methods to classify automatically interword spaces into boundaries or non-boundaries –which comes down to segment into sentence-level units– while trying to figure out which features are the most efficient to identify boundaries. We thus built six different segmenters which we present here.

2. Macrosyntactic Units for Spoken Language

The segmentation we are aiming at is based on the units defined in the Rhapsodie project and described in this section. However, the number and heterogeneity of the boundaries to retrieve (up to 18) heavily complicates the task. We therefore decided to train the segmenters on a simplified version.

2.1. Rhapsodie Syntactic Annotation

In the reference treebank, we distinguished two levels of syntactic cohesion: the first one, called microsyntax, is ensured by government relations and is the traditional level of syntax considered by most syntactic theories and treebanks. The second one, called macrosyntax, is the cohesion of sequences, Illocutionary Unit (henceforth IU) which are wrapped up into a unique illocutionary act, realising one and only one assertion, injunction, interrogation, etc. (Blanche-Benveniste et al., 1990; Berrendonner, 1990; Cresti, 2000; Deulofeu et al., 2010).
The text is thus segmented into IUs noted by // as in the following example:

> *ˆdonc quand vous arrivez sur la place de Verdun <+ la Préfecture est sur votre droite // vous < vous continuez tout droit // en fait < "euh" "ben" là < "euh" vous passez "euh" & // "ben" en fait < c'est la ligne du tram > toujours > là //*
> [Rhap-M0001, Avanzi corpus]

> ˆso when you arrive on place de Verdun <+ the prefecture is on your right // you < you continue straight away // in fact < "er" "well" there < you pass "er" & // "well" in fact < it's the tram line > still > there //

An IU may be composed of several segments: the main segment, called the nucleus, bears the illocutionary force; separated from the nucleus by < and > are respectively the pre-nuclei (such as *donc quand vous arrivez sur la place de Verdun* 'so when you arrive on place de Verdun') and the post-nuclei ; IU introducers (*donc* 'so') are marked by the symbol ˆ while discourse markers (*ben* 'well', *euh* 'er') are surrounded by double quotation marks " ".

The macrosyntactic annotation of Rhapsodie has been carried out by a dozen of scholars working cumulatively. Each part of the corpus has been checked by at least three annotators and in case of strong disagreement, group sessions have been conducted to discuss the most problematic items. Apart from those debates, no evaluation of the inter-annotator agreement has been done.

2.2. Simplified Annotation for the Task

In order to reduce the complexity of the task, differents types of boundaries have been grouped together. As a

result, in the experiments we present, boundaries between IUs are considered major boundaries (//) while those around segments of IUs are considered minor boundaries (noted / for all kinds of segments). Each interword space can have one and only one value : two similar boundaries are merged into one, and major boundaries are prioritised in the event of a conflict with minor boundaries.

The previous example would then be annotated:

> *donc / quand vous arrivez sur la place de Verdun / la Préfecture est sur votre droite // vous / vous continuez tout droit // en fait / euh / ben / là / euh / vous passez / euh // ben / en fait / c'est la ligne du tram / toujours / là //* [Rhap-M0001, Avanzi corpus]

> so when you arrive on place de Verdun / the prefecture is on your right // you / you continue straight away // in fact / er / well / there / you pass / er // well / in fact / it's the tram line / still / there //

We can roughly consider those boundaries respectively as dots and commas in written texts. To give an idea of the distribution of boundaries in the reference corpus, only 10% of interword spaces are major boundaries, 13% are minor boundaries and 77% are not macrosyntactic boundaries. This disproportion causes a bias we took into account in the evaluation of the experiments.

3. Features

As the treebank we worked on was provided with rich annotations, we could use a large set of features. However, although this data is interesting in terms of determining segmentation cues and may be reliable enough to guarantee a good outcome, it is not likely to be available in the corpora that are intended as inputs for macrosyntactic segmenters and we therefore felt the need to use more objective information by means of state-of-the-art tools. This alternative also ensures the possibility of a completely automatic segmentation task.
Thus, among all the features described in this section, some were directly taken from the manual contribution of naive and expert annotators which was subsequently validated by linguistics experts. On the other hand, a few more features had to be extracted from TextGrids (the preferred format of transcribed speech, produced via the Praat program (Boersma, 2002)) and others were calculated either using Analor[1], a semi-automatic annotation tool for prosodic structures, or using SEM (Tellier et al., 2012), a morphosyntactic tagger for French built based on a Conditional Random Field (CRF) algorithm. At the end of the process, using a data architecture specifically built to work at the interface of syntax and prosody, parser Farouch (Belião and Liutkus, 2014) enables all data to be gathered together in a single table and properly aligned on tokens. The collection of all those cues is shown in Figure 1.

In order to provide classifiers as much information as possible, all features span six positions (three words or three syllables before and after each interword space). To determine the class of an interword space, segmenters are then given the possibility to access and use cues such as whether or not the preceding word is a determiner, as well as whether there is a change of speaker after the first, second or third following word.

3.1. Prosodic Features

Two types of prosodic features were available for our experiments: acoustic and perceptual data.

Acoustic characteristics we considered as interesting cues were (1) the pitch (both the raw value and the average value calculated on a span of two syllables before and after the current one), (2) syllable and pause duration, (3) the rising rate of the pitch and (4) the slowing coefficient : a normalised measure computed by Gendrot et al. (2012) based on the comparison between the duration of each syllable in the corpus and a typical reference duration. A high coefficient indicates the significant lengthening of a syllable and should not only help identify specific events such as hesitation or emphasis but also help detect the end of syntactic or prosodic units, commonly associated with a rhyme lengthening due to a final accent in spoken French.
No speaker normalisation has been applied to acoustic values.

Along with those features, we also used perceptual notions: (5) intonative periods defined by Lacheret and Victorri (2002) as maximal intonational units used in the description of prosodic discourse organisation and characterised by their register (low, normal, high) and by their shape (falling, flat, rising), (6) the strength of prominences (or salient syllables), (7) hesitation, and (8) global contour and local register levels, describing the pitch level compared to the average range of the speaker as a contour or a single value for a given unit. Those four features are available in the Rhapsodie treebank, but periods and prominences were also automatically calculated by Analor (Avanzi et al., 2008; Avanzi et al., 2011).

3.2. Morphosyntactic Features

The corpus was also automatically (9) part-of-speech (POS) tagged and (10) shallowly parsed into chunks.

Token	POS	Chunk
je	CLS	B-NP
suis	V	B-VN
née	VPP	I-VN
à	P	B-PP
Cannes	NC	I-PP
pendant	P	B-PP
la	DET	I-PP
guerre	NC	I-PP

Figure 2: Example of morphosyntactic tagging with SEM on an excerpt of [Rhap-D2004, Lacheret corpus]

Both taggings were successively performed by the same

[1] http://www.lattice.cnrs.fr/Analor?lang=en

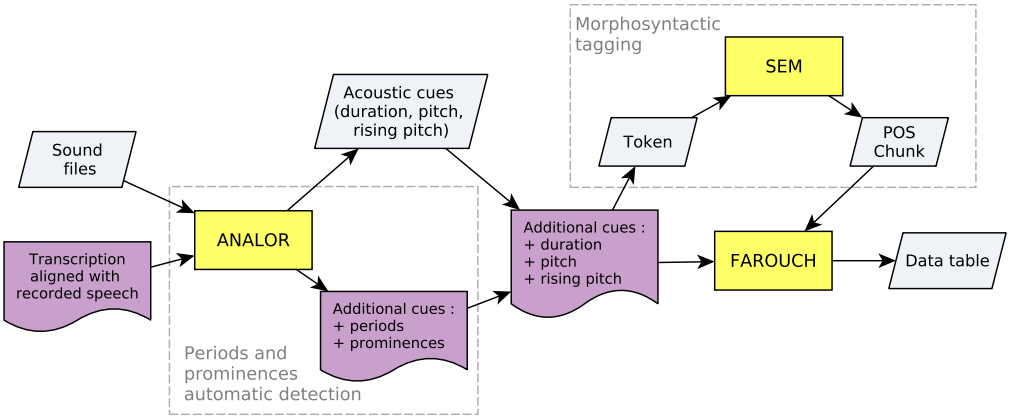

Figure 1: Flowchart of data pre-processing in the collection of prosodic and morphosyntactic features

tool which was trained on written data but although the margin of error when SEM[2] is applied to oral data is relatively significant (around 17% loss on the accuracy score for POS tagging, and a similar impact on chunking), learning a whole new set of tags adapted to transcribed data has a heavily important cost in time and effort. However, an adapted chunker is easier to train and future experiments could use the same approach as the third one described in Tellier et al. (2013).
Tags are produced as shown in Figure 2.

According to Abney (1991), a chunk is the "non-recursive core of an intra-clausal constituent, extending from its beginning to its head". Here, the definition of a chunk extends to pre-head phrases as well but stays within the limits of non-recursive phrases, with chunks being the smallest possible. In the example, the sentence is chunked [I] [was born] [in Cannes] [during the war] using two tags from the BIO (Beginning, In, Out) encoding.

3.3. Word Features

Our set of features includes other cues such as (11) speaker diarisation (which can be helpful in dialogs, and especially in overlapping sequences) and (12) speech disfluency. The former was directly extracted from one of the layer of annotation in the Rhapsodie treebank and for that reason, cannot be used in a fully automated segmentation, while the latter simply uses an exhaustive list composed of discursive markers or words that are specific to oral data (for instance, *euh* 'uh', *bon* 'well', or backchannel *ouais* 'yeah', etc.). In addition, if a word is cut off in mid-utterance (most of the time transcribed with a hyphen or a tilde at the end), the phenomenon is to be noted with the value of the speech disfluency feature being in this case *amorce* (false start).

[2]The tool is freely available on the following page, along with a description and the annotation guide used : www.lattice.cnrs.fr/sites/itellier/SEM.htm

4. Experiments

Supervised classification experiments were conducted using the open source software Weka (Hall et al., 2009), a data mining tool developed by the University of Waikato. Different categories of algorithms from the variety implemented in Weka were tested (notably NaiveBayes, Rules and Decision Trees) but only results from the J48 algorithm are presented in this paper as it proved to be the best performing on our data set. As with any decision tree algorithm, is possible to read and visualise the models we learned.
Using all the features at once could lead to the best results but in order to verify this assumption, we separated the features considering their nature and the way they were obtained - either from a manual annotation or an automated process. This setting also allows us to compare the efficiency of each type of feature for the task.

Six different input files were therefore created by selecting and filtering the features we wanted to be taken into account in each model:

1. All features (mixed): with all the features described above (from 1 to 12).
2. All features (auto): with all the features we could collect automatically, either prosodic (1-6) or part of the morphosyntactic and word features (9, 10, 12).
3. Prosody only (mixed): with all the prosodic features (1-8).
4. Prosody only (auto): with automatically calculated prosodic features only (1-6).
5. No prosody (mixed): with all but the prosodic features (9-12).
6. No prosody (auto): with all automatically calculated features, except for the prosodic ones (9, 10, 12).

Experiments presented in this paper were limited to classification into three classes only: major syntactic boundaries, minor syntactic boundaries, or non-boundaries. A known

problem in classifying such data is the bias caused by non-boundaries outnumbering actual boundaries. As a consequence, a simple algorithm with no other rule than classifying all instances in the majority class makes a high baseline. In a previous study (Wang, 2013), an attempt to address the bias was made by under-sampling non-boundaries but this failed to improve the performance of classifiers.

5. Results and Evaluation

In this section we present the results of the models we trained using a 10-fold cross-validation method.

Accuracy	F-measure	Weighted F-measure
77.4%	0.87	0.68

Table 2: Majority class baseline results

As expected, the baseline for these experiments (Table 2) is quite high for the proportion of non-boundaries makes a 77,4% accuracy baseline but with only 0.68 overall weighted F-measure. This score was obtained with the algorithm ZeroR provided by Weka, which simply puts every instance in the most common class (in our case, that means classifying each interword space as non-boundaries).

The first observation we can make by comparing it to the results of our models (Table 1) is that all of the experiments we conducted have better results, gaining up to 7% accuracy for the best performing models.
What then draws our attention with a closer look at Table 1 is the similarity between the scores of (a) models using all features and (b) models using the automatic ones only, with a margin as slim as 0.1% accuracy, which is not a significant difference considering the size of our data set. Furthermore, using all features or excluding prosodic cues also makes almost no difference. It just seems that "no prosody" models have better precision scores for both major and minor boundaries, whereas models using all types of features have the best recall scores which is quite important because the main problem for classifiers seems to be the coverage of all boundaries, not the precision. As for classifications based on prosodic cues only, they unexpectedly do merely better than the baseline.
As a matter of fact, the Rhapsodie corpus gathers 53 samples of about five minutes each for a total amount of three hours of speech recording. The objective was to maximise the variety of speakers and speech situations. The resulting heterogeneity might have interfered in the classification process by making it difficult for algorithms to identify reliable values for features, especially for prosodic features. Indeed, as shown in the study of Belião (2013) on this corpus, the ratio between IUs and intonative periods may vary highly according to the situation of utterance: in political speeches, for instance, the speaker tends to cut IUs into several periods, while in narrative sequences such as those found in interviews, the speaker is inclined to utter several IUs within a single period. Finding a generic patt is not possible with those two contradictory patterns
We can eventually notice that the lowest scores are systematically related to minor boundaries with F-measure scores never reaching 0.5 (especially due to low recall), questioning the uniformity of those boundaries and the choice that was made to group them together to simplify the task. Further analysis of the results and experiments will help determine whether minor boundaries should be split back into different groups or not, and in that case, how they should be categorised.

6. Conclusion

Macrosyntatic segmentation is the first step to linguistic analyses such as parsing. Many works have studied the use of machine learning methods for speech segmentation on English corpora and have proved successful (Stolcke and Shriberg, 1996; Favre et al., 2008). The study we present in this paper is the application of similar techniques but on French data with similar results, besides the unexpected low scores of models using prosodic cues only. However, this outcome needs to be put into perspective as the corpus we used for these experiments contains different types of speeches (monologs, dialogs, interactive or

(a) Both automatic and manual features

Model	Annot.	P	R	F1	Acc.
All	//	0.62	**0.55**	**0.58**	84.3%
	/	0.6	**0.35**	0.44	
	none	**0.89**	0.96	**0.92**	
Prosody only	//	0.57	0.39	0.46	79.2%
	/	0.36	0.08	0.16	
	none	0.82	0.96	0.89	
No prosody	//	**0.64**	0.49	0.55	**84.5%**
	/	**0.68**	0.31	**0.43**	
	none	0.87	**0.98**	**0.92**	

(b) Automatic features

Model	Annot.	P	R	F1	Acc.
All	//	0.61	**0.55**	**0.58**	**84.4%**
	/	0.61	**0.36**	**0.45**	
	none	**0.89**	0.96	**0.93**	
Prosody only	//	0.51	0.37	0.43	77.8%
	/	0.27	0.09	0.14	
	none	0.83	0.94	0.88	
No prosody	//	**0.63**	0.47	0.54	84.3%
	/	**0.68**	0.31	0.42	
	none	0.87	**0.98**	0.92	

Table 1: Cross-validation results in terms of information retrieval metrics : Precision, Recall, F-measure and Accuracy

semi-interactive speeches etc.) characterised by prosodic specificities as we mentioned in the previous section. To address this issue, future experiments will test a new feature to identify automatically the type of speeches by running a preliminary clustering task. Prosodic features should be more valuable cues if combined with that information. This process would then be prepended to the main workflow if it proves to be efficient.

Another noteworthy point in building not only one segmenter but different segmenters is that it facilitates application on new data. The whole process is illustrated in Figure 3, where only one "yes" to one of the three options (acoustic cues and/or morphosyntactic cues and/or manual annotation if available) is necessary to start classifying. Furthermore, Weka offers the possibility to select features used in training models making this method fully adaptable to the available features of a new corpus.
We can also note that with only three features, namely discursive markers, POS and chunking tags (automatic features without prosody so no alignment is required), it is possible to achieve 84.3% accuracy, whereas the best model using all features including manual annotation does only 0.2% better. Considering the efforts put into manual annotation and the slight improvement in the outcome of the best model, the minimal segmenter might be an interesting candidate to process non-tagged corpora.

As part of the Orfeo project, these segmenters are meant to be used as a preprocess of newly acquired large corpora and to accelerate the work of syntactic annotators. Experiments are currently conducted on both tagged and untagged versions of the TCOF corpus (Benzitoun et al., 2012) and will soon be evaluated to determine whether the errors made by the segmenters are significant, and whether the use of fully automatic segmenters is indeed time saving when combined to an additional correction phase and worthwhile over pure manual segmentation. Further examination of decision trees will also help measure the discriminating power of features.

The macrosyntactic segmenters presented in this paper are meant to be open source tools and will thus soon be freely available online[3]. Work is still in progress so please contact authors for further information on the state of advancement of the resource.

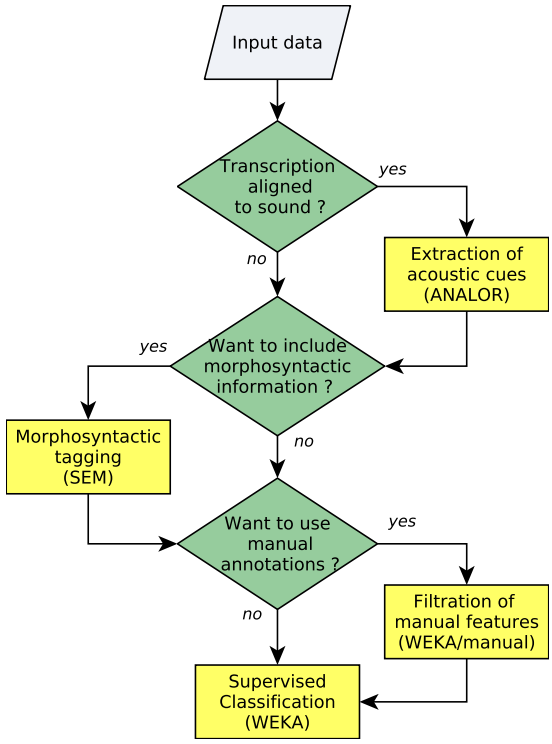

Figure 3: Process flowchart of a new corpus to be segmented depending on its nature and on the level of information it includes

7. Acknowledgements

This study was made possible by the support of the ANR through funding of the Orfeo project.

8. References

Abney, S., (1991). *Parsing by Chunks*. Kluwer Academic Publishers, Dordrecht, R. Berwick, S. Abney and C. Tenny (eds.) edition.

Avanzi, M., Lacheret-Dujour, A., and Victorri, B. (2008). Analor. A tool for semi-automatic annotation of French prosodic structure. *Proceedings of Speech Prosody 2008*.

Avanzi, M., Obin, N., Lacheret, A., and Victorri, B. (2011). Toward a continuous modeling of French prosodic structure: Using acoustic features to predict prominence location and prominence degree. In *Proceedings of Interspeech*, pages 2033–2036.

Belião, J. and Liutkus, A. (2014). Farouch : une architecture de données pour l'analyse d'interfaces linguistiques. In *Congrès Mondial de Linguistique Française 2014. CMLF'2014*.

Belião, J. (2013). Characterizing oratory speech through both prosody and syntax. *Preproceedings of the ESSLLI'2013 Student Session*, 1(18):1–12.

Benzitoun, C., Fort, K., and Sagot, B. (2012). TCOF-POS: un corpus libre de français parlé annoté en morphosyntaxe. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 99–112.

Berrendonner, A. (1990). Pour une macro-syntaxe. *Travaux de linguistique*, (21):25–36.

Blanche-Benveniste, C., Bilger, M., Rouget, C., Van Den Eynde, K., Mertens, P., and Willems, D. (1990). *Le français parlé. Etudes grammaticales*. Editions du CNRS, Paris.

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345.

Cresti, E. (2000). *Corpus di italiano parlato*. Accademia della Crusca, Florence.

Deulofeu, J., Duffort, L., Gerdes, K., Kahane, S., and Pietrandrea, P. (2010). Depends on what the French say spoken corpus annotation with and beyond syntactic

[3]A link to the project will be posted on this page http://www.lattice.cnrs.fr/sites/itellier/TEOK.html.

functions. In *Proceedings of the Fourth Linguistic Annotation Workshop. LAW IV*, pages 274–281.

Favre, B., Hakkani-Tür, D., Petrov, S., and Klein, D. (2008). Efficient sentence segmentation using syntactic features. In *Spoken Language Technology Workshop 2008. SLT'2008*, pages 77–80.

Gendrot, C., Adda-Decker, M., and Schmid, C. (2012). Comparaison de parole journalistique et de parole spontanée : analyses de séquences entre pauses. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1 : JEP*, pages 649–656, Grenoble, France.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Lacheret, A. and Victorri, B. (2002). La période intonative comme unité d'analyse pour l'étude du français parlé: modélisation prosodique et enjeux linguistiques. *Verbum*, 1(24):55–72.

Lacheret, A., Kahane, S., Belião, J., Dister, A., Gerdes, K., Goldman, J.-P., Obin, N., Pietrandrea, P., and Tchobanov, A. (2014). Rhapsodie: a Prosodic-Syntactic Treebank for Spoken French. In *Proceedings of the 9th Language Resources and Evaluation Conference, LREC'2014*.

Pietrandrea, P., Lacheret, A., Kahane, S., and Sabio, F. (2014 (to appear)). The notion of sentence and other discourse units in spoken corpus annotation. In Mello, H. and Raso, T., editors, *Spoken corpora and Linguistic Studies*. John Benjamins Amsterdam.

Stolcke, A. and Shriberg, E. (1996). Automatic linguistic segmentation of conversational speech. In *Proceedings of the Fourth International Conference on Spoken Language 1996. ICSLP'96*, volume 2, pages 1005–1008.

Tellier, I., Dupont, Y., and Courmet, A. (2012). Un segmenteur-étiqueteur et un chunker pour le français. *Actes de TALN'12, session démo*.

Tellier, I., Dupont, Y., Eshkol, I., and Wang, I. (2013). Adapt a text-oriented chunker for oral data: How much manual effort is necessary? In *Intelligent Data Engineering and Automated Learning–IDEAL 2013*, pages 226–233. Springer.

Wang, I. (2013). Segmentation automatique d'un corpus de français oral en unité macrosyntaxiques. Master's thesis, Université Paris 3.