Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6389–6394
Marseille, 11–16 May 2020

Automatic Period Segmentation of Oral French

Natalia Kalashnikova[1,4], Loïc Grobol[2], Iris Eshkol-Taravella[3], François Delafontaine[1]

LLL UMR 7270[1], Lattice[2], MoDyCo UMR 7114[3], ZAION[4]

10 Rue de Tours, 45065 Orléans , 1 Rue Maurice Arnoux, 92120 Montrouge, 200 Avenue de la République 401B, 92001 Nanterre, 18bis Rue de Villiers, 92300 Levallois-Perret
natalia.kalashnikova@sorbonne-nouvelle.fr, loic.grobol@ens.fr, ieshkolt@parisnanterre.fr, francois.delafontaine@etu.univ-orleans.fr

Abstract

Natural Language Processing in oral speech segmentation is still looking for a minimal unit to analyze. In this work, we present a comparison of two automatic segmentation methods of macro-syntactic periods which allows to take into account syntactic and prosodic components of speech. We compare the performances of an existing tool Analor (Avanzi, Lacheret-Dujour, Victorri, 2008) developed for automatic segmentation of prosodic periods and of CRF models relying on syntactic and / or prosodic features. We find that Analor tends to divide speech into smaller segments and that CRF models detect larger segments rather than macro-syntactic periods. However, in general CRF models perform better results than Analor in terms of F-measure.

Keywords: spoken language, automatic segmentation, period, oral french, CRF, macro-syntactic units

1. Introduction

Automatic segmentation of oral speech is one of the main components of spoken language processing applications such as speech recognition, data extraction, etc. However, sentences, the relevant units for written language are ill-suited for these tasks. Lacheret and Victorri (2002) suggest to replace it by the notion of *prosodic period*, grounded in observations and analyses of spontaneous spoken language. A semi-automatic tool for speech segmentation in periods, Analor, was developed (Avanzi, Lacheret-Dujour, Victorri, 2008) within this theory. However, this approach *only* takes into account prosodic characteristics of speech, which – since segmentation depends not only on prosody, but also on syntax and semantics – seems overly restrictive.

We propose to go beyond these limitations by working within the Fribourg model of macro-syntax, which defines periods in both syntactic and prosodic terms. In order to develop an automatic tool capable of detecting macro-syntactic periods, we cast this segmentation task as a sequence labelling problem and study the adequacy of machine-learning models using Conditional Random Fields (Lafferty et al. 2001) relying on lexical, syntactic and prosodic features.

This work is a part of the SegCor[1] project whose goal is to develop several tools for automatic segmentation of linguistic units, including *periods* which will be our focus. The result of one of these tools for the segmentation of chunks (Eshkol-Taravella et al., 2019) is used in this study.

The article is presented as follows. Section 2 represents the related work of periods ; section 3 describes the corpus and goals of this study ; section 4 illustrates the main points of the manual annotation that serves as a reference to test the automatic methods; section 5 presents experiments of automatic annotations ; the results are analyzed in section 6 ; and section 7 proposes a conclusion and some perspectives for further work.

2. Related works

This study is based on the difference between two approaches of periods. Lacheret and Victorri (2002) define a period as a prosodic structure which connects several syntactic constructions within one discursive block. In this case the aim of periods is to make speech more coherent. One syntactic structure can also be organized within several prosodic periods. Thus, periods act as a way of topicalization. The end of a period is marked with a pause, which in French is of at least 300 milliseconds :

(1) et vous logez euh () la le la façade
du théâtre (0.72)
and you continue euh () the the the facade of the theater (0.72)

In example (1) the end of the period is detected after the word « théâtre » because of the length of the pause. Analor (Avanzi, Lacheret-Dujour, Victorri, 2008) is a semi-automatic segmentation tool developed within this framework.

The concurrent approach of the Groupe de Fribourg (2012) considers periods as an autonomous prosodic units defined by their conclusive intonational shape (Berrendonner, 2017). Macro-syntactic approaches rely on prosody to analyze the syntactic structure of a spoken language (Blanche-Benveniste et al., 1990; Cresti et al., 2011). For that purpose the period potentially constitutes both a complete structure and a maximal monologic unit (Groupe de Fribourg 2012: 34-35). There is no tool for automatic segmentation of macro-syntactic periods.

In this study we aim to determine the most performant method of automatic segmentation for *macro-syntactic* periods. To achieve this goal, we compare two methods. The first is Analor which does not involve the training process and does not analyze the syntax of periods. The second a learning segmentation considered as a labeling task as in (Eshkol-Taravella et al., 2019; Tellier et al., 2012, 2013, 2014) using CRF models. This method allows to take into account syntactic features as well as prosodic.

[1] SegCor : http://segcor.cnrs.fr/

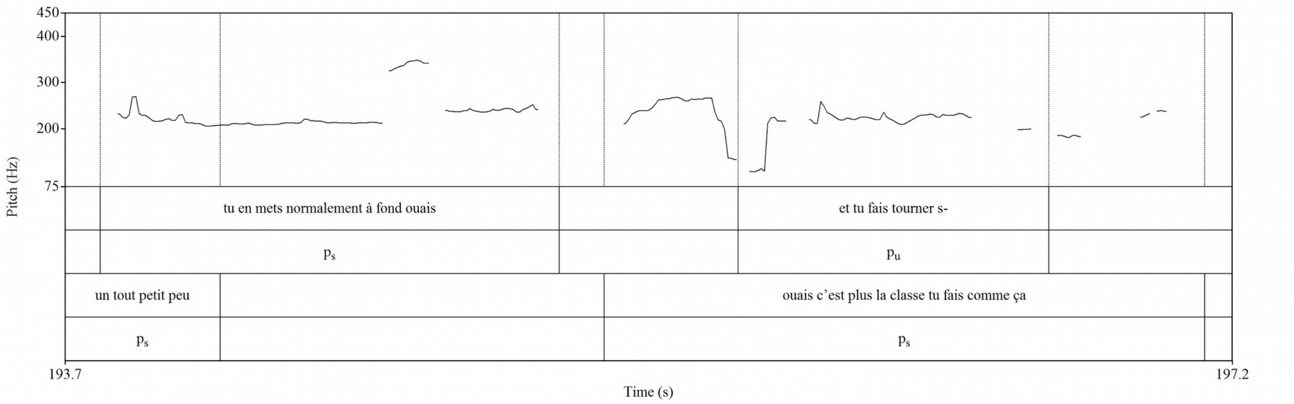


Figure 1: Intonative pattern

3. Corpus and Goals

This study is realized within the SegCor project. The project aims to study the oral segmentation in French and German. The French part of the project is constructed of the ESLO2 (Baude & Dugua 2011, Eshkol et al., 2011) and CLAPI (Balthasar & Bert, 2005) corpora. ESLO2 and CLAPI corpora contain recordings of spontaneous speech in different situations. The recordings are anonymised and transcribed for research use.

For our study we work on a subcorpus of 10 dialogue transcriptions of 10 minutes each and 1 monologue transcription of 20 minutes. The extracts represent different types of speech in terms of conversational environment, relationships between the speakers, etc. Thus, this pilot corpus contains transcriptions of preparing a meal, meetings, conferences, radio transmissions, interviews, etc. The pilot corpus is manually annotated in pragma-syntactic periods (see Section 4).

This study compares two methods of automatic annotation of macro-syntactic periods. The first one uses Analor, which detects periods using only prosodic hints. This tool comes with default settings which should perform well on French but can be tuned (see Section 5). Since Analor uses only prosodic features, is it enough to detect macro-syntactic periods or do we need another method that takes into account the syntax and other prosodic features? The second method relies on CRF sequence labelling models, which have been shown to perform well for segmentation tasks (Eshkol-Taravella et al. 2019; Tellier et al., 2012, 2013, 2014). The performance of both methods is compared to manual annotation of the pilot corpus.

4. Manual Annotation

The manual annotation (by one annotator in Praat (Boersma and Weenink, 2001) relied on the definition of *period* from the Fribourg macro-syntactic model (Groupe de Fribourg, 2012). One of two criteria needed to be met for the manual annotation: either the detection of a conclusive intonational contour or an effective change of a speaker.

The detection of conclusive contours was done perceptively, alongside with the speaker listening to the recordings. While the model provides conclusive contour patterns (Groupe de Fribourg, 2012: 109), the annotation relied on the same properties as with Anne-Lacheret's framework used here for the automatic annotation. The Fribourg model, in practice, long relied on that framework (Avanzi, 2005, 2012), and the list of conclusive contour patterns proved hard to apply.

As for a change of a speaker, it covered either the speaker interrupting his or her speech due to another speaker intervening, or the speaker abandoning an incomplete structure after a lengthy pause (more than 0.8 second). In the latter case, it must be distinguished from the speaker revising his or her structure without such a pause.

From those guidelines two types of units could be annotated, either complete (p_s) or incomplete (p_u):

2) ELI tu en mets normalement à fond ouais (0.5)
you usually fill it yeah (0.5)

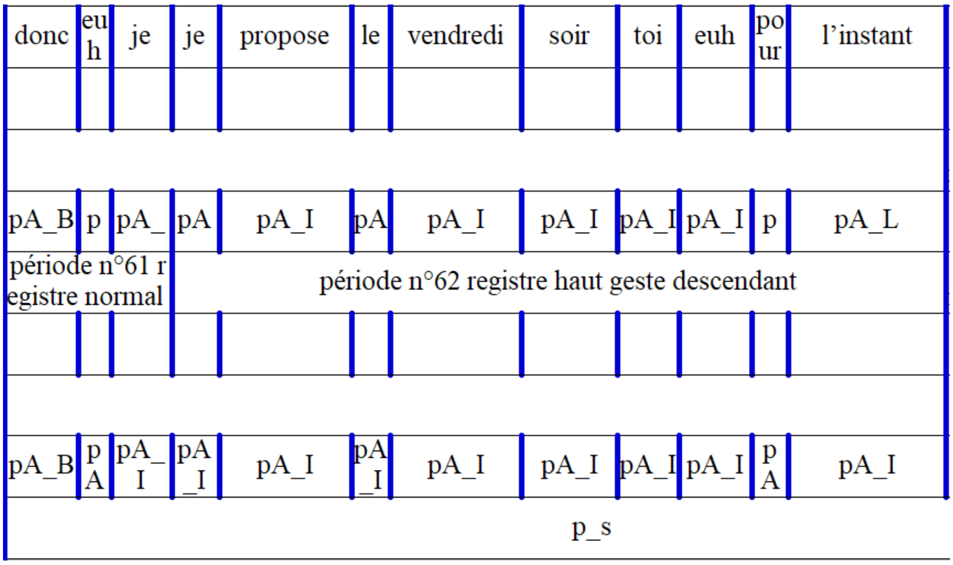
BEA ouais [c'est plus classe tu fais] comme ça
yeah [it's classier you do] like that

ELI [et tu fais tourner]
[and you turn it o-]

Figure 1 illustrates example (2) and completes it with a pitch. ELI's first period is followed by a speaker change, alongside a perceived rising conclusive contour that the acoustic pitch fails to detect. ELI's second period is interrupted due to the overlap, without any conclusive contour (but a lengthy pause afterward). BEA's period as reported in example (2) showcases a structure with multiple micro-syntactic units.

Beyond speaker change, syntax also proved to play a role for the handling of pauses, which can either be a simple suspension of the speaker's speech, or the speaker abandoning his or her turn even momentarily, for example to request help for lexical completion (Lerner, 1991). In a monologic context, this distinction becomes critical, including for pauses under 0.8 second. To resolve such cases, reliance on prosodic cues alone proved insufficient. Beyond the weighing of prosodic properties, syntactic completion was the decisive factor:

3) ELI elle a fait son master sa première année de
master (0.7) ah de psy cho



donc	euh	je	je	propose	le	vendredi	soir	toi	euh	pour	l'instant
pA_B	p	pA_	pA	pA_I	pA	pA_I	pA_I	pA_I	pA_I	p	pA_L
période n°61 registre normal			période n°62 registre haut geste descendant								
pA_B	pA	pA_I	pA_I	pA_I	pA_I	pA_I	pA_I	pA_I	pA_I	pA	pA_I
p_s											

Figure 2: Example of TextGrid.

The first line represents tokens of the sequence, the fourth and the eighth contain the same intervals as tokens but which are annotated as a part of periods, the fifth - periods detected by Analor and the last manually annotated periods.

she worked on a master a first year of
master (0.7) ah in psycho

In example (2), intonation is continuative (with a flat contour) before the 0.7 second pause, with no other speaker intervention. While the speaker is looking for lexical completion (psycho), the continuative contour (despite pause length) and a micro-syntactic structure bridging that pause led to the rejection of a periodic boundary here.

Other remarks on the guidelines, such as the handling of speaker change with the first speaker continuing, would exceed the purpose of comparing manual and automatic annotations, as they do not affect the resulting units.

5. Experiments

5.1 Analor

The corpus is annotated with the semi-automatic processing tool Analor (Avanzi, Lacheret-Dujour, Victorri, 2008). Segmentation of periods applies 4 criteria: 1) pause lasting for at least 300 milliseconds; 2) difference in height between the mean value of fundamental frequency over all the signal before the pause and the last value of fundamental frequency before the pause; 3) difference in height between the last value of fundamental frequency before the pause and the first one after the pause; 4) absence of hesitation (« euh ») just before or after the pause. Thus, Analor considers period as a prosodic segment between two pauses with its own melodic shape. Taking into account the difference of periods definitions, performance of the software for annotation of macro-syntactic periods is worth exploring.

5.1.1 Preprocessing

Analor requires Praat format files a PitchTier and a TextGrid to launch a procedure of annotation. PitchTier is a file that contains the shape and the values of the sound pitch. We establish the scale of fundamental frequency for each audio file and then use it as a setting to create a PitchTier. TextGrid (as in Figure 2) contains 3 tiers: word (one interval per token of recording), speaker (one interval per speaking slot) and manual annotation of periods (one interval per period for each speaker).

5.1.2 Experiments

Analor is a pre-trained tool so we can use the whole data set for segmentation. After launching the procedure of annotation, Analor creates another TextGrid file for every sound file containing a new tier of automatically segmented periods. However, Analor creates only one tier with periods but TextGrid files of manual annotation contain a tier for each speaker in every sound file. We solve this problem by manually dividing this tier into one tier per speaker. Another problem is that the number of periods of automatic annotation is not the same as that of the manual. The solution is to tokenize automatic and manual annotated periods using BILU tagging. BILU is the abbreviation, where B stands for beginning (the first element of the sequence), I - in (for elements between the first and the last ones), L - last (the last element) and U - unique (the only element in sequence).

5.2 CRF Models

Linear chain Conditional Random Fields (CRF, Lafferty et al. 2001) are sequence labelling models designed for efficient machine-learning and robustness to long-distance effects. CRF take into account contextual information from previous labels to make predictions. In particular, CRF models have been shown to be well-suited for the formulation of segmentation as sequence labelling that we use here for other tasks such as chunking (Eshkol-Taravella et al., 2019; Tellier et al., 2012, 2013, 2014) or Named Entity Recognition (Dupont & Tellier 2014).

5.2.1 Features

Our CRF models are developed using two kinds of features: prosodic and morpho-syntactic. Prosodic features are the values of fundamental frequency (minimum, maximum and mean), of intensity (minimum, maximum and mean) and duration of each token. Morpho-syntactic features are POS tags labeled by TreeTagger (Schmids, 2014) and Chunk tags developed by (Eshkol-Taravella et al., 2019) for the SegCor project. Thus, we

word	f0 min	f0 mean	f0 max	duration	int min	int mean	int max	BILU
ça	9	9	10	82	39	40	41	pA_B
va	7	8	9	83	39	41	43	pA_L
dis	6	9	14	82	40	41	41	pA_B
je	10	13	14	81	41	42	43	pA_I
voulais	7	8	10	81	42	42	43	pA_I
te	6	7	9	80	42	43	43	pA_I
de- mander	6	9	10	86	42	42	42	pA_L
demain	6	9	14	84	38	39	40	pA_B

Table 1: Prosodic features for the CRF models

« pA_B » stands for the beginning of a period, « pA_L » - the end and « pA_I » is a label for tokens inside the period.

develop three types of CRF models: the first one is built only on prosodic features, the second one – on prosodic and morpho-syntactic and the third one - only on one of prosodic features. We do it with the aim to answer the question: do prosodic features contain enough information to realize segmentation of Fribourg macro-syntactic periods or do we need morpho-syntactic features as well?

5.2.2 Preprocessing

During the pre-processing phase acoustic features are extracted for each token using Praat. Prosodic values are divided into groups of values to facilitate the training of CRF models. The values of intensity are divided by 10, of fundamental frequency by 20 and of duration by 0.1. TextGrid files containing tokens and morpho-syntactic features are transformed into tables. We collect all data to create two types of tables for training: the first with prosodic features (P corpus) and the second with prosodic and morpho-syntactic features (P+M corpus). Table 1 shows a sample of a data set of a period with only prosodic features.

The data is organized by sequences of utterances, so one utterance can contain several periods. Taking into account that the initial pilot corpus does not have a great amount of data and the number of utterances containing several periods within it, we decide to extend the corpus by multiplying the existing data. We save the same values of features but replace tokens by placeholder words. It allows a system to have more similar data of values for training without memorizing the values of existing words.

We also analyze the influence of intensity and fundamental frequency by training models that use only one of these prosodic features. The result of the performance of models based only on one prosodic feature can show which one is more important for period's segmentation.

The corpus is divided into 3 sets: 60 % train, 30 % test and 10 % development. In total we have 6 different configurations for training CRF models: 1) initial corpus with prosodic features (I_1P corpus), 2) initial corpus with prosodic and morpho-syntactic features (I_1P+M corpus), 3)initial corpus with only morpho-syntactic features (I_1M corpus), 4) initial corpus with fundamental frequency's features (I_2F0 corpus), 5) initial corpus with intensity's values (I_2INT corpus), 6) augmented corpus with only prosodic features (E_1P corpus).

5.2.3 Experiments

CRF models are developed using the Wapiti software (Lavergne, Cappé, Yvon, 2010). It is a toolkit that uses different discriminative models for segmenting and labelling sequences.

In total, we developed 6 different models to establish the most performant combination of features and pre-processing procedure.

6. Results

In table 2, we report the performance of both methods analyzed in terms of precision, recall and F-measure of the periods deduced from the BILU labels of the automatically and manually annotated periods. In cases, when Analor detects a period that was not annotated manually and vice versa we use a label "pA_O" where "O" stands for "out".

For Analor, the score of precision is higher than the score of recall. It is due to the fact that Analor detects smaller segments than the macro-syntactic periods. Moreover, in most cases the highest results correspond to speakers with least time of speaking and conversely.

For the CRF models, it seems that using only morphosyntactic features already yields better results than using the periods detected by Analor, they are less useful

Model	P	R	F
Analor	0.52	0.22	0.31
Prosody	0.56	0.78	0.66
Morphosyntax	0.54	0.68	0.60
Prosody + morphosyntax	0.56	0.78	0.66
Prosody + morphosyntax + augmentation	0.56	0.70	0.62
F0	0.55	0.76	0.64
Intensity	0.72	0.55	0.62

Table 2: Comparisons of the different methods

than purely prosodic features, and combining them does not yield further improvements.

To define the importance of each of prosodic features we compare the results of models built only on the values of fundamental frequency and only on the values of intensity. The results show a great complementarity between these features, with each contributing to a different aspect of detection and their association obtaining better results than any of the single features.

7. Conclusions and further work

In this paper we presented a new method for automatic segmentation of oral speech. We analyzed macro-syntactic periods which allow to take into account syntactic and prosodic content of speech.

Analor's results show that the tool detects periods from prosodic perspective without analyzing pragma-syntax. The performance of the software is not satisfying enough for the annotation of Fribourg macro-syntactic periods.

All of the CRF models get better scores than Analor. The F-measure varies from 0.54 to 0.66 among different CRF models. If we compare the performance of every CRF models and take into account time of pre-processing the most performant model is developed on an initial corpus with the first scale of values.

We also defined than the model based only on fundamental frequency features showed better results than the model developed with intensity features. To find out the importance of different prosodic characteristics on the performance of CRF models could be one of the ways for further research. Another way of pre-processing data for CRF models could be rather the difference between the values of current word with the previous.

Taking into account the better results of CRF models we can conclude that the difference in period's definitions of two approaches is crucial for automatic segmentation.

8. Acknowledgements

This work was supported by the ANR project SegCor (ANR-15-FRAL-0004). We also thank the lab Lattice and the organization ZAION for providing support for this research.

9. Bibliographical References

Avanzi, M. (2005). Quelques hypothèses à propos de la structuration interne des périodes.

Avanzi, M., Lacheret, A., Victorri, B. (2008). Analor, un outil d'aide pour la modélisation de l'interface prosodie-grammaire. CERLICO, France. pp.27-46.

Avanzi, M. (2012). L'Interface prosodie/syntaxe en français. Dislocations, incises et asyndètes. Bruxelles: Peter Lang.

Balthasar, L. & Bert, M. (2005). « La plateforme « Corpus de langues parlées en interaction » (CLAPI) », Lidil, 31, 13-33.

Baude, O. & Dugua, C. (2011) (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ? In *Corpus*, Varia, 10, pp.99-118.

Berrendonner, A. (2017). La notion de période (note terminologique) in *Encyclopédie grammaticale du français.*

Blanche-Benveniste, C., Bilger, M., Rouget, C., Eynde (van den), K. (1990). Le français parlé. In *Études grammaticales*. Paris: CNRS.

Boersma, P. & Weenink, D. (2019). Praat: doing phonetics by computer [Computer program]. Version 6.1.07, retrieved 26 November 2019 from http://www.praat.org

Cresti, E., Moneglia, M., Tucci, I. (2011). Annotation de l'entretien d'Anita Musso selon la théorie de la langue en acte. In *Langue française*, 170: 95-110.

Dupont, Y. & Tellier, I. (2014) A Named Entity recognizer for French Proceedings of TALN 2014 (Volume 3: System Demonstrations).

Eshkol-Taravella I., Baude O., Maurel D., Hriba I., Dugua C., Tellier I. (2011) Un grand corpus oral « disponible » : le corpus d'Orléans 1968-2012. in *TAL: Ressources linguistiques libres*. Volume 53, n° 2, p. 17-46

Eshkol-Taravella I., Maarouf M., Skrovec, M., Badin, F. (2019). Chunker différents types de discours oraux : défis pour l'apprentissage automatique Proceedings of TALN 2019 , 1-5 juillet 2019, Toulouse, France.

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

Lafferty, J., McCallum, A., Pereira, F. (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of ICML.

Lacheret-Dujour, A. & Victorri, B. (2002) La période intonative comme unité d'analyse pour l'étude du français parlé : modélisation prosodique et enjeux linguistiques. In M. Charolles (ed), *Verbum* Nancy, pp. 55-72, 2002.

Lerner, G.H. (1991). On the syntax of sentences-in-progress. In *Language in Society*, 20: 441-458.

Schmid, H. (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

Tellier, I., Duchier, D., Eshkol, I, Coumet, A., Martinet, M. (2012). Apprentissage automatique d'un chunker pour le français. Proceedings of TALN 2012, Grenoble, France

Tellier I., Dupont Y., Eshkol I., Wang I. (2013). Adapt a Text-Oriented Chunker for Oral Data: How Much Manual Effort is Necessary?, The 14th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'2013), Special Session on Text Data Learning, Hefei, China.

Tellier I., Eshkol-Taravella, I., Dupont, Y., Wang, I.. (2014). Peut-on bien chunker avec de mauvaises étiquettes POS ? Proceedings of TALN 2014, Marseille, France.