10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Speaking Style Prosodic Variation and the Prosody-Syntax Interface
A Large-Scale Corpus Study

George Christodoulides[1]

[1]Metrology and Language Sciences Unit, Université de Mons, Belgium
george@mycontent.gr

Abstract

As large spoken language corpora become available, we revisit previous analyses based on smaller datasets and verify whether the conclusions generalise to the new data. In this paper, we present an analysis of speaking style variation in French, based on a large-scale corpus (450 hours, 2500 speakers), and compare it with previous analyses that were based on smaller corpora. The corpus is segmented at the phonetic, syllabic and word level; automated annotation in parts-of-speech and syntactic dependencies was performed, enhancing existing annotations; and a multitude of acoustic and prosodic features are automatically extracted. Statistical analysis (clustering, PCA) is performed to explore the characteristics of speaking styles, individual variation, and the discriminatory power of different sets of prosodic and linguistic features. We also endeavour to model the relationship between prosodic units and syntactic units, on various levels of granularity, and we explore congruency and mismatch between these units as a method to discriminate speaking styles.

Index Terms: speaking style, prosodic variation, classification and clustering of speaking styles, prosody-syntax interface

1. Introduction

Situational variation, i.e. the linguistic variation related to differences in communicative situations, is important for the study of language and speech. The systematic study of such variation, in the crossroads between sociolinguistics and phonetics/phonology, is the domain of socio-phonetics. Speaking style variation is being actively explored, and the dichotomy between ”laboratory speech” and ”natural speech” has been questioned [1]. Speaking style is determined by the situational context and by individual characteristics ([2], [3], [4]). The situational context is best described using multiple dimensions (cf. the model proposed in [5]) rather than binary distinctions (e.g. ”formal” vs. ”informal”).

The effects of speaking style on the prosodic features of speech are studied through corpus-based studies (e.g. [6], [7]), sometimes with a view to a specific application such as the automatic classification of speaking styles (e.g. [8], [9]), the description of atypical speech (e.g. [10]), or in an attempt to characterise the dynamics of conversational interaction (e.g. [11], [12]).

As larger spoken language corpora become available, it is possible to revisit previous analyses that were based on smaller corpora, in order to verify whether the findings generalise to the larger dataset, and to perform more fine-grained analyses. This paper presents a study on a large corpus of spoken French (see section 2.1) and its methodology is inspired from [6]. We also present findings related to the relationship between prosodic and syntactic segmentation, across speaking styles, based on the same corpus.

2. Methodology

2.1. Corpus

The present study was performed on the spoken part of the *Corpus d’Etude pour le Français Contemporain* (CEFC) corpus [13]. This is a collection of French spoken language corpora from various sources ([14], [15], [16], [17], [18], [19]) that have been homogenised, transcribed, annotated automatically for parts-of-speech and dependency syntax, and have been approximately aligned to the word level. The composition of the corpus is presented in Table 1. The corpus contains 900 samples, has a duration of 300 hours, and covers more than 2500 speakers. Its length is approximately 3.7 million tokens, which after phonetisation correspond to approximately 4.6 million syllables.

Speaking Style	Nb	Dur	Spk	Syll	Tok
Private					
Activity	10	2,3	14	21,4	27,4
Conversation	174	65,5	488	902,5	1075,2
Dining	12	8,0	39	102,0	120,7
Interview	351	137,2	829	1672,3	2076,4
Narration	37	8,3	43	89,4	111,7
Public					
Activity	13	8,1	126	62,7	77,9
Conversation	81	6,2	209	67,6	86,5
Interview	9	3,2	31	42,1	52,6
Lesson	16	4,3	63	35,3	49,2
Media	31	10,4	271	117,2	163,0
Meeting	50	28,1	319	348,6	443,6
Narration	86	15,8	96	148,2	188,3
Public speech	30	5,9	66	58,9	84,3
Total	900	303	2594	3668,4	4556,9

Table 1: *Composition of the CEFC corpus: number of samples, duration (in hours), number of speakers, syllables (in thousands) and tokens (in thousands).*

Due to the fact that the CEFC corpus has been composed from various sources, a large variety of communicative situations are represented. In order to organise the samples, the meta-data of the CEFC corpus use four main dimensions: domain (public or private), type/genre (providing a broad description of the activity or the communicative situation), context (e.g. friendly, family, business, political, academic, etc.) and channel (in public, face to face, radio, TV, telephone). A distinction is also made between monologues, dialogues and multi-party conversations. We have combined the domain attribute with the type/genre attribute (grouping some communicative situations that are under-represented), in order to arrive at the categorisation in speaking styles presented in Table 1. In this paper, we present our results (mainly) across this grouping of *genres* or *speaking styles*; however further analyses can be performed on the more detailed *sub-genres*.

10.21437/SpeechProsody.2020-144

2.2. Data Processing

Initially, and in order to improve the performance of the automatic text-speech alignment, a restoration and enhancement procedure was applied to all audio samples of the corpus, using *iZotope RX 6 Audio Editor*. The following filters were applied in sequence: de-clip (restore clipped samples at high quality), de-click (remove random clicks), de-hum (remove hum noise and harmonics), voice de-noise (adaptive noise reduction), equaliser match (using the “full dialogue” preset) and leveller (normalisation of audio levels, respecting dialogue dynamics). The corpus meta-data (TEI-encoded XML files) and annotations (CoNLL-U files with speaker information and the approximate alignment time codes) were all imported into an SQL database using the corpus management software *Praaline* [20] for further processing.

A phonetic transcription with pronunciation variants was produced from the orthographic transcription, and was aligned at the phone, syllable, word and utterance levels, using *Praaline*’s Forced Aligner; for French, it uses *Kaldi* [21] for speech recognition and a pronunciation lexicon based on GLÀFF [22].

The corpus was re-annotated using *DisMo* [23] producing part-of-speech tags and an automatic detection of disfluencies and multi-word units; this annotation was combined with the original annotation in dependency syntax. The aligned corpus was analysed using *ProsoGram* [24], which detects the vocalic nucleus of each syllable based on its intensity and voicing; the F0 curve is then stylised into a pitch curve (of static and dynamic tones) based on a perceptually-grounded approach. We subsequently applied the plug-in *Promise* in order to perform an automatic detection of prosodically prominent syllables [25] and an automatic detection of major and minor prosodic boundaries [26]; the statistical algorithms of *Promise* have been trained on manually annotated French corpora, as detailed in the tool’s references. An automatic segmentation in prosodic units (intonation phrases and accentual phrases) has been performed on the basis of these annotations, and is correlated to the syntactical annotation as outlined in section 3.4.

We have finally used the statistical analysis tools in *Praaline* (Temporal Analysis, Prosodic Profile, and Units) in order to extract multiple acoustic and prosodic features (measures) for each corpus sample and each speaker’s participation in each sample. These measures can be grouped as follows: temporal measures (e.g. pause duration, speech rate); conversational dynamics measures (length of turns, gaps and overlaps); pitch measures (e.g. pitch register and dynamic movements); prominence measures; and unit-related measures (prosodic units, syntactic units and their correlation). The database from Praaline was linked to the R statistical software [27] for analysis.

3. Results

In the following, we present some statistically significant results based on the *genre* classification, for each group of prosodic descriptors.

3.1. Temporal Features

The articulation ratio (percentage of sample time when a speaker is articulating speech) per speaking style is plotted in Figure 1. The ”priv-narration” style (spontaneous narration of personal experiences), as well as the ”prof-narration” style (professionally narrated fairy tales) have lower ratio, indicating longer silent pause time. A similar observation can be made for the ”prof-lesson” style (academic lectures and school lessons), as well as the ”prof-public speech” style (where pauses are mainly used for rhetorical effect). The results with respect to articulation rate (articulated syllables per second) are presented in Figure 2: we also observe that professional narration has a lower rate, while the differences between the other speaking styles remain minor (variability is also higher in spontaneous conversations).

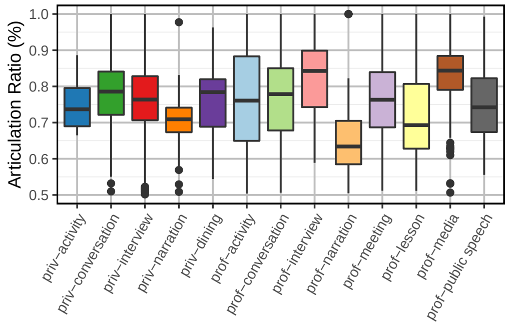

Figure 1: *Articulation ratio (%) per speaking style.*

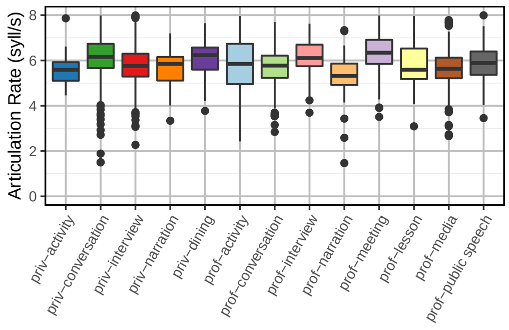

Figure 2: *Articulation rate (syll/s) per speaking style.*

The number of filled pauses (normalised to the number of tokens) is presented in Figure 3 and is a prosodic descriptor that discriminates those speaking styles (and sub-genres) that have a low degree of planning, or no previous planning at all (e.g. conversation, whether in a private or professional setting, interviews). However, we observe that filled pauses (and the results are similar for other types of disfluencies) are present in all speaking styles (with the exception of pre-planned professional narration).

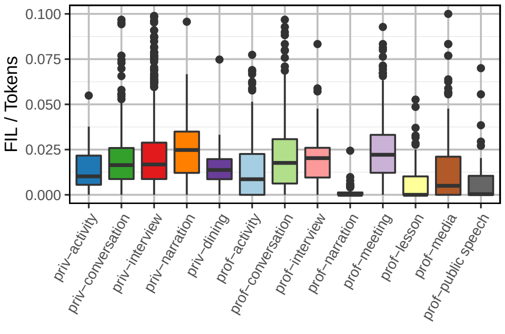

Figure 3: *Filled pauses (normalised to the number of tokens) per speaking style.*

3.2. Conversational Dynamics

With respect to conversational dynamics, Figure 4 show the mean duration (in seconds) of a turn of speech. The differences between more and less interactive speaking styles can thus be observed, especially if the turn duration is combined with the percentage of overlaps and gaps (i.e. silences between turns of different speakers), across styles, shown in Figure 5.

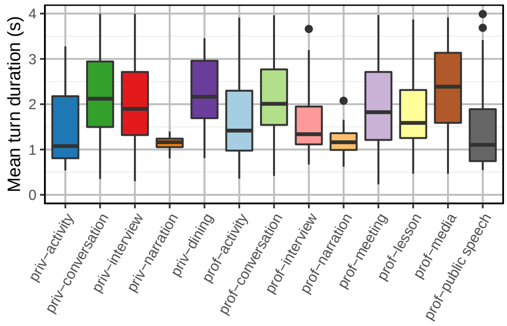

Figure 4: *Mean turn duration (s) per speaking style.*

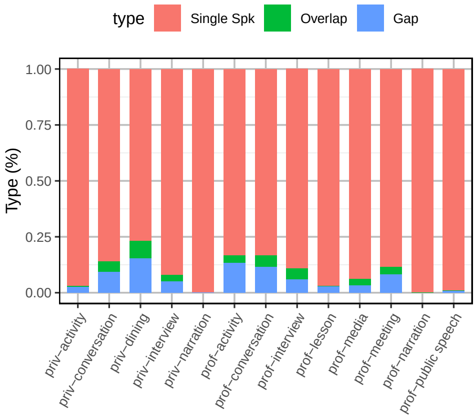

Figure 5: *Conversational dynamics per speaking style.*

3.3. Intonation

Figure 6 shows the distribution of pitch trajectory per channel of communication (this measure is calculated as the sum of absolute pitch intervals within syllabic nuclei, divided by duration, and it is measured in semitones per second). We observe that the two media-related activities (radio and TV) have larger trajectories, which can be explained from the adoption of a more expressive style by professional presenters.

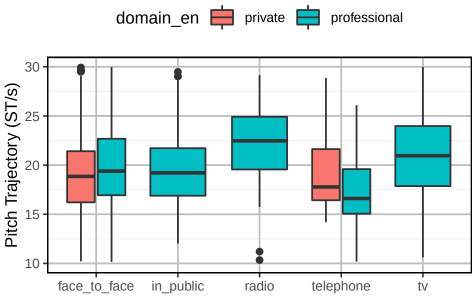

Figure 6: *Pitch trajectory (ST/s) per channel.*

3.4. Prosody-Syntax Interface

The CEFC corpus is annotated for syntax, using a simplified dependency syntax framework. The syntactic annotation of spoken language presents difficulties and choices have to be made regarding the treatment of sentence fragments, parentheticals and disfluencies. For that reason, in the present study we have used the provided annotation and limited the analyses to the relationship between major syntactical segments (i.e. complete dependency units) and prosodic segments. The dependency annotation defines ”sentence-like units” (SU) and their size in syllables, across speaking styles, is shown in Figure 7.

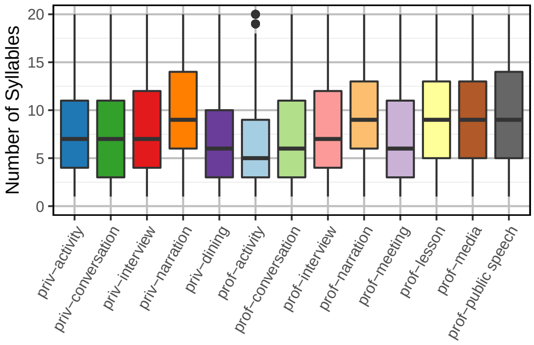

Figure 7: *Number of syllables per sentence unit, per speaking styl˜*

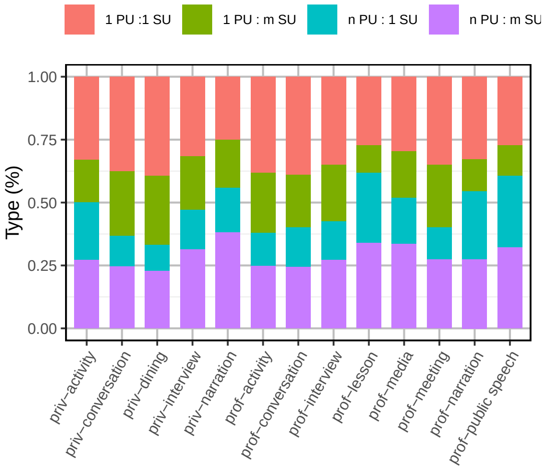

Figure 8: *Relationship between prosodic and syntactic units, per speaking style.*

Based on the automatic detection of prosodic boundaries and prominent syllables, we annotate the corpus in prosodic units (PU): intonation units and accentual units. There are congruences and mismatches between these units, giving four possible configurations: one PU corresponding to one SU (1:1), one PU spanning multiple SUs (1:m), multiple PUs produced for one SU (n:1) and a series of mismatches of prosodic/syntactical boundaries that leads to multiple PUs corresponding to multiple SUs (n:m). This method of analysis is similar to the ones previously presented by [7] and [28] for smaller multi-genre corpora of French. The results of the analysis of the congruence and mismatch between major syntactical and prosodic units, across speaking styles, are shown in Figure 8. As speaking styles are characterised by differences in speech planning, these affect the length of syntactic structures, and the number and length of silent pauses (the main acoustic correlate of major prosodic boundaries); we can therefore use the metrics of congruence/mismatch of units in order to differentiate between styles.

3.5. Principal Component Analysis

We have confirmed that no unique prosodic parameter is sufficient to differentiate between speaking styles. Given that many of the extracted measures are highly correlated, we proceed by applying a Principal Component Analysis, that reduces the set of measure to an small set of linearly uncorrelated *principal components* (each principal component is a linear combination of the initial measures). The PCA results indicate that the first 2 principal components (PCs) explain 25.1% of the variance, the first 4 PCs explain 44.6% of the variance, and 8 PCs explain 71.0% of the variance. Compare these results with the ones reported in [6] (with 9 speaking styles and 105 samples), where the first 2 PCs explained only 43% of the variance and the first 8 explained 78.2%.

Figure 9 shows a plot of the individual points (each representing the participation of a speaker in a corpus sample), colour-coded by speaking style, and plotted with the first 2 principal components (PC1 on the x axis and PC2 on the y axis). Confidence ellipses (plotted in Figure 10) indicate the variability of each speaking style: we can thus observe that some speaking styles are highly homogeneous (e.g. media presentations), while conversations and sociolinguistic interviews have a higher variability.

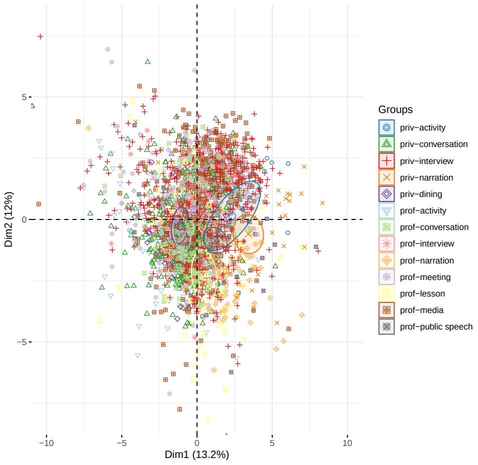

Figure 9: *First two principal components and all samples per speaking style.*

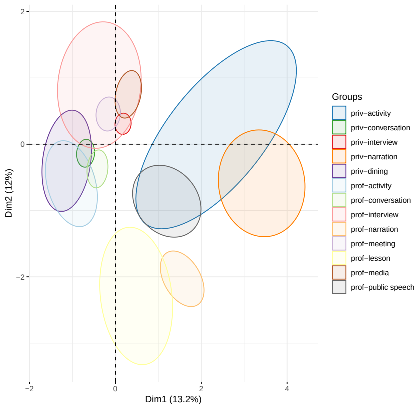

Figure 10: *First two principal components and confidence ellipses per speaking style.*

By analysing the contribution of each measure in each PCA dimension, we observe that PC1 is related to interactivity (turn changes, overlaps), PC2 captures global pitch characteristics (e.g. pitch range), PC3 captures the degree of expressiveness (e.g. pitch trajectories, rises and falls) and PC4 captures the temporal characteristics (articulation ratio, median duration of silent pauses and speech rate).

4. Conclusion and Perspectives

In this article we have presented an exploratory descriptive analysis of the prosodic variation of speaking styles in a large-scale corpus (300 hours, more than 2500 speakers) of French. While the general tendencies of previous studies on smaller corpora are largely confirmed, the analysis on these ”big data” indicates that a much more nuanced approach to speaking style variation is necessary: some communicative situations have very specific constraints (e.g. professional speech in the media, or political public speaking, but also impromptu narrations of personal experiences) that influence multiple prosodic parameters and are thus easier to classify; however, the individual variation in spontaneous conversation, or even sociolinguistic interviews is high. This finding from a big-corpus study reinforces the calls for even more diversity in speech research: an effort must be made not only to study prosodic phenomena across speaking styles, but also to embrace individual variation in the analyses.

In a study similar to the present one, Ryant and Liberman [9] attempt to select a set of prosodic measures that can be automatically extracted and used for a multi-dimensional characterisation of a large variety of speech datasets, in English, Spanish and Chinese. Provided that both the data and the processing tools are open and freely available, it can be envisaged to combine these studies, in order to gain insight into cross-linguistic differences.

In future studies, using the enhanced version of the CEFC corpus, we envisage a more detailed analysis of the prosody-syntax interface (exploring multiple levels of constituents inside sentence units); an exploration of alternative methods to describe speaking styles (different dimensions); and extending the extracted features (especially in the segmental level). Finally, our work has enriched and improved the annotation of a large corpus of spoken French (the corpus is available under the Creative Commons license, and our work is made available under the same conditions).

5. References

[1] P. Wagner, J. Trouvain, and F. Zimmerer, "In defense of stylistic diversity in speech research," *Journal of Phonetics*, vol. 48, pp. 1–12, 2015.

[2] J. Llisterri, "Speaking styles in speech research," pp. 15–17, 08 1992.

[3] M. Eskenazi, "Trends in speaking styles research," in *Eurospeech*, 1993.

[4] P. Léon, *Précis de phonostylistique, Parole et expressivité*. Paris: Nathan Université, 1993.

[5] P. Koch and W. Österreicher, "Language of immediacy – language of distance: Orality and literacy from perspective of language theory and linguistic history," in *Communicative spaces: Variation, contact, and change*, C. Lange, B. Weber, and G. Wolf, Eds. Frankfurt: Peter Lang, 1985/2012, p. 441–473.

[6] J.-P. Goldman, T. Pršir, G. Christodoulides, and A. Auchlin, "Speaking style prosodic variation: an 8-hour 9-style corpus study," in *7th International Conference on Speech Prosody, May 20–23, Dublin, Ireland, Proceedings*, 2014, pp. 105–109.

[7] J. Beliao, S. Kahane, and A. Lacheret, "Modéliser l'interface intonosyntaxique," in *Prosody-Discourse Interface Conference 2013, Proceedings*, 10 2013.

[8] A. Veiga, D. Celorico, J. Proença, S. Candeias, and F. Perdigão, "Prosodic and phonetic features for speaking styles classification and detection," in *Advances in Speech and Language Technologies for Iberian Languages. Communications in Computer and Information Science*, D. Torre Toledano, Ed. Berlin, Heidelberg: Springer, 2012, vol. 328, pp. 15–17.

[9] N. Ryant and M. Liberman, "Automatic analysis of phonetic speech style dimensions," in *Interspeech*, 2016, pp. 77–81.

[10] K. Hudry, C. Aldred, S. Wigham, J. Green, K. Leadbitter, K. Temple, K. Barlow, H. McConachie, P. Consortium *et al.*, "Predictors of parent–child interaction style in dyads with autism," *Research in Developmental Disabilities*, vol. 34, no. 10, pp. 3400–3410, 2013.

[11] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

[12] J. Grothendieck, A. L. Gorin, and N. M. Borges, "Social correlates of turn-taking style," *Computer Speech and Language*, vol. 25, pp. 789–801, 2011.

[13] C. Benzitoun, J.-M. Debaisieux, and H.-J. Deulofeu, "Le projet ORFÉO : un corpus d'étude pour le français contemporain," *Corpus*, vol. 15, pp. 91–114, 2016, Actes du colloque Corpus de Français Parlés et Français Parlés des Corpus.

[14] S. Branca-Rosoff, S. Fleury, F. Lefeuvre, and M. Pires, *Discours sur la ville. Présentation du Corpus de Français Parlé Parisien des années 2000 (CFPP2000)*, 2012. [Online]. Available: http://cfpp2000.univ-paris3.fr

[15] E. Cresti, F. Bacelar do Nascimento, A. Moreno Sandoval, J. Veronis, P. Martin, and C. Kalid, *The C-ORAL-ROM Corpus. A Multilingual Resource of Spontaneous Speech for Romance Languages*. John Benjamins Publishing Company, 2005.

[16] H. Baldauf-Quilliatre, I. Colon de Carvajal, C. Etienne, E. Jouin-Chardon, S. Teston-Bonnard, and V. Traverso, "CLAPI, une base de données multimodale pour la parole en interaction : apports et dilemmes," *Corpus*, vol. 15, pp. 165–194, 2016, Actes du colloque Corpus de Français Parlés et Français Parlés des Corpus.

[17] M. Avanzi, M.-J. Béguelin, and F. Diémoz, "De l'archive de parole au corpus de référence. Le corpus oral de français de Suisse romande (OFROM)," *Corpus*, vol. 15, pp. 309–342, 2016, Actes du colloque Corpus de Français Parlés et Français Parlés des Corpus.

[18] J. Carruthers, "Annotating an oral corpus using the Text Encoding Initiative: Methodology, problems, solutions," *Journal of French Language Studies*, vol. 18, no. 1, pp. 103–119, 2008.

[19] Équipe Delic, *Autour du Corpus de référence du français parlé*. Publications de l'université de Provence, 2004, Recherches sur le français parlé No 18, 265 pp.

[20] G. Christodoulides, "Praaline: integrating tools for speech corpus research," in *LREC 2014 – 9th International Conference on Language Resources and Evaluation, May 26–31, Reykjavik, Iceland, Proceedings*, 2014, pp. 31–34. [Online]. Available: http://www.praaline.org

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, IEEE Catalog No.: CFP11SRW-USB.

[22] N. Hathout, F. Sajous, and B. Calderone, "GLÀFF, a Large Versatile French Lexicon," in *LREC 2014 – 9th International Conference on Language Resources and Evaluation, May 26–31, Reykjavik, Iceland, Proceedings*, 2014.

[23] G. Christodoulides and G. Barreca, "Expériences sur l'analyse morphosyntaxique des corpus oraux avec l'annotateur multi-niveaux DisMo," *Corela: Cognition, Représentation, Langage*, vol. HS-21, 2017, journals.openedition.org/corela/4867.

[24] P. Mertens, "The Prosogram: Semi-automatic transcription of prosody based on a tonal perception model," in *Proc. of Speech Prosody 2004, March 23–26, Nara, Japan*, 2004, pp. 549–552.

[25] G. Christodoulides and M. Avanzi, "An evaluation of machine learning methods for prominence detection in French," in *Interspeech 2014 – 15th Annual Conference of the International Speech Communication Association, September 14–18, Singapore, Proceedings*, 2014, pp. 116–119.

[26] G. Christodoulides, "Acoustic correlates of prosodic boundaries in French: A review of corpus data," *Revista de Estudos da Linguagem, Belo Horizonte*, vol. 26, no. 4, pp. 1531–1549, 2018, aop13597.2018.

[27] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017. [Online]. Available: https://www.R-project.org/

[28] L. Martin, L. Degand, and A. Simon, "Forme et fonction de la périphérie gauche dans un corpus oral multigenres annoté," *Corpus*, vol. 13, pp. 243—-265, 2014.