C-PROM-Task: A New Annotated Dataset for the Study of French Speech Prosody

Mathieu Avanzi[1], *Lucie Rousier-Vercruyssen*[1], *Sandra Schwab*[2], *Sylvia Gonzalez*[1], *Marion Fossard*[1]

[1] Chaire de logopédie, University of Neuchâtel, Ruelle Vaucher 22, Neuchâtel, 2000, Switzerland
[2] ELCF, University of Geneva, Rue Candolle 5, Geneva, 1211, Switzerland
mathieu.avanzi@unine.ch, lucie.rousier-vercruyssen@unine.ch, sandra.schwab@unige.ch, sylvia.gonzalez@unine.ch, marion.fossard@unine.ch

Abstract

The aim of this paper is to describe C-PROM-Task, a dataset created and annotated in the same spirit as C-PROM [1]. C-PROM-Task was annotated following a perceptually-based and computer-assisted procedure for the study of syllabic prominences, syllabic disfluences and two ranges of prosodic units. All told, C-PROM-Task comprises recordings and annotated TextGrids of story-telling by 20 native French speakers from Switzerland. The entire dataset is 2 hours 20 minutes long. Some observations are also made regarding accentuation (prominence rate), disfluence rate and phrasing (length of prosodic units) in the corpus.

Index Terms: corpus, spoken French, prosodic annotation, prominence, phrasing.

1. Introduction

Until fairly recently, annotation of continuous French speech was either relatively rudimentary and approximate or the province of a group of specialists working within the framework of phonologic theories. On the one hand, specialists of spoken French ([2], [3] and [4]) transcribe prosodic events using a reduced set of symbols, which does not reflect the actual complexity of prosodic phenomena. On the other hand, phonologists who work in the Autosegmental-Metrical (AM) framework [5], such as [6] and [7], use or develop annotation systems which are not really applicable to spontaneous speech since the data they deal with mostly consist of laboratory speech, that is "light years ahead of the complexity of spontaneous speech" [8]. However, in the past few years, mostly thanks to automatic processing advances, the situation has been changing. Protocols and tools designed to annotate French prosodic structure (semi-)automatically are emerging (see [9] for an overview). An increasing number of projects aim to create available and public annotated corpora [10]. In this context, the purpose of this paper is not to make an inventory of existing systems and resources but (i) to present C-PROM-Task, a prosodically annotated corpus based on the same hypotheses and with the same aims as C-PROM [1]; (ii) to summarize the perceptually-based and computer-assisted procedure used to annotate accentuation (calculation of the position and strength of pitch accents within a given group of words) and phrasing (identification of the different prosodic groups in the prosodic hierarchy) in this corpus; and (iii) to briefly discuss what such annotations can teach us about French speech prosody.

2. Method

2.1. Participants and task

Twenty French-speakers from Switzerland (10 male and 10 female) took part in the study. Ten of the participants were from 19 to 27 years old (mean age: 22.6, SD: 2.2), and the other ten were between 71 and 82 years old (mean age: 75.5, SD: 3.1). We will refer to these two groups as the "young group" and the "older group". The participants in the two groups were strictly matched for gender. Speakers were recorded in a storytelling situation. They were asked to describe verbally 18 different story picture sequences with 3 increasing levels of complexity according to the number and gender of the characters involved in the story picture sequence, i.e., level 1 or easy level: one character; level 2 or medium level: 2 characters of different genders; level 3 or difficult level: 2 characters of the same gender. Half of the stories were presented with a logical order of events (logical condition) and the other half with a non-logical order (non-logical condition). For each of the three complexity levels and each of the two conditions (logical vs. non-logical order), speakers had to describe three different story picture sequences. Using a referential communication paradigm (see [11] and [12]), the storytelling in sequence test enables one to assess how a participant (the director of the interaction) plans his/her discourse and what type of verbally discriminating information he/she produces that will enable an addressee (the researcher) to identify and order the 6 pictures that constitute a story sequence. To avoid non-verbal communication, the participant and researcher were separated by an opaque screen. For each interaction, stories were presented in a pseudo-randomized order and verbal productions were recorded.

2.2. Selection of the files

We processed one story for each level of complexity for both orders, making six stories in all for each speaker. Thus the C-PROM-Task corpus contains 120 files (3 levels of difficulty*2 order conditions*20 speakers). The total duration of the corpus is 2 hours 20 minutes.

2.3. Annotations

2.3.1. Orthographic transcription and text-to-sound alignment

Each of the 120 files was first orthographically transcribed within Praat software [13]. Transcriptions were then semi-automatically aligned in phones, syllables and words with

EasyAlign [14] script. Alignments were manually checked and corrected when necessary by two of the authors (each author was in charge of half of the data). Silent pauses and non-transcribed segments (interventions from the researcher, overlapped speech, laughs, etc.) were transcribed with the symbol "_".

2.3.2. Annotation of syllabic prominences and disfluencies

Syllabic prominences and disfluencies were manually annotated in parallel by two of the authors, using the method described in [1]. To summarize, the annotators had to listen to small stretches of the signal (2-3 seconds on average), 3 times at most, and to code in a dedicated tier (an empty copy of the syllable tier) with "p" and "P" the syllables they perceived as weakly and strongly prominent. They were asked to annotate the syllables perceived as associated with a disfluency with "H" (false starts, breaks in the syntactic program, elongations due to a hesitation, "euh", etc.). To ensure the coding was performed on perceptual bases as far as possible, the researchers did not have visual access to acoustic information (f0 and intensity lines, spectral envelope). The Analor tool [15] was then used to obtain an automatic annotation of prominent syllables in a new tier. The algorithm calculates the relative height and duration of each syllable in a given stretch of speech by comparing the value of the analyzed syllable with the average of the six adjacent syllables (i.e. three preceding and three following ones); the pitch rise slope is then processed and the presence of a subsequent silent pause is considered. Thresholds to activate prominence were the ones trained for spontaneous speech indicated in [15], that is to say 1.5 for relative duration, 1.3 st for relative height and 2.5 st for melodic rise. To avoid false-alarm prominence detection, [16]'s algorithm was used to ensure that pitch path files were as clean as possible.

The inter-annotator agreement coding was statistically tested. Regarding prominence, "p" and "P" were merged and considered together as a single category (which contrasts with "0", see [1] for the justification). The total number of intervals considered was 18'604 (syllables associated with an "H" were not taken into account). First, Cohen's kappa [17] was used to assess reliability between a pair of annotators. It appeared that the agreement between the two human annotators was substantial ($\kappa = 0.68$), while it was fair between the first annotator and Analor ($\kappa = 0.56$) and between the second annotator and Analor ($\kappa = 0.48$). Fleiss' Kappa, a measure used to assess reliability between more than two annotators [19], indicated a fair agreement between the three annotators ($\kappa = 0.57$). Regarding disfluences, Cohen's kappa revealed an almost perfect agreement ($\kappa = 0.82$) out of the 12'530 intervals taken into account (syllables annotated with the symbols "p" or "P" were excluded from the calculation). A syllable was considered prominent in the reference tier if it was marked in two of the three annotation tiers. The final status of a syllable hesitating between "H" and something else was decided after discussion between the two annotators.

2.3.3. AP segmentation

Next, a tier indicating the boundaries of minor prosodic units was obtained as follows: each final syllable of a lexical word or polysyllabic functional word that was coded as prominent generated the boundary of a minor prosodic constituent, including every element without prominence on its left side. Following the AM theory [20], we will refer to these units as Accentual Phrases (henceforth APs), even though in many cases they are closer to Clitic Groups than to APs. When the last syllable's item was a schwa, the penultimate syllable was considered as carrying the final pitch accent, thus marking the right boundary of the AP. Syllables labeled "H" are either comprised in the AP of the surrounding valid syllables or form an AP on their own (such syllables are excluded from the calculations presented below).

2.3.4. IP segmentation

In spite of its importance for speech processing, the definition of the major prosodic units called Intonational Phrases (IP) in the AM framework is still an issue for scholars working on French [17]. It is either described in terms of syntactic/information structure (root clauses, embedded coordinated clauses, left- and right-peripheral constituents map onto IPs) or regarding their intonational realization: IPs are defined by the presence of a nuclear accent (syllable associated with a major pitch movement, a pre-boundary lengthening and/or followed by a silent pause). To identify IP in the C-PROM-Task database, the prominence degree detection function provided by the Analor tool [15] was used. On the basis of four automatically measured acoustic parameters (relative syllabic duration, relative f0 average, slope contour amplitude and presence of an adjacent silent pause), the software estimates a degree of strength for the last syllable of each AP on a scale from 0 to 10 (from the least to the most prominent). The calculations rely on two fundamental principles. The first is a quantity principle: the greater the number of acoustic parameters involved in the identification of a prominence and the distance from predetermined thresholds, the stronger the prominence is perceived. The second is a compensation principle, which stipulates that if one of the classic parameters involved in the perception of prominence in French presents a low value and another presents a high value, there will be the same feeling of prominence as if the two parameters involved both presented a medium score. We considered that the last syllable of an AP was associated with a nuclear contour, i.e. an IP boundary, if its strength reached a score of 4/10.

3. Analysis

We illustrate and briefly comment on some descriptive statistics regarding the distribution of the syllables in the corpus according to their status (+/- prominent, +/- disfluent). We then discuss the effects of different factors influencing accentuation and phrasing.

3.1. Descriptive Analysis

3.1.1. Distribution of the syllables according to their labels

Among the 21'161 syllabic intervals in our corpus, 7'546 were identified as prominent (35.65%), 2'464 as disfluent (11.64%) and 11'151 were left blank (52.69%), as seen in Figure 1:

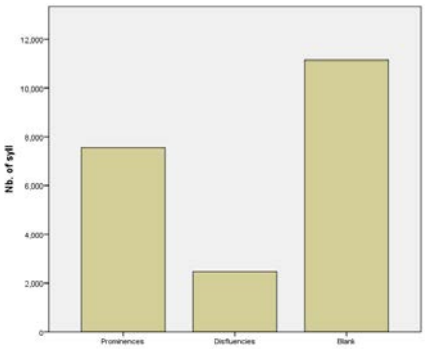

Figure 1. Distribution of the syllables in the corpus according to their type. From left to right: prominent syllables, disfluent syllables and blank syllables.

3.1.2. *AP length*

In total, the corpus contains 6'547 APs. The mean length of an AP is 2.85 syllables. On average, as seen in Figure 2 below, an AP mostly comprises between 2 and 4 syllables (80.3% of the data), rarely less or more.

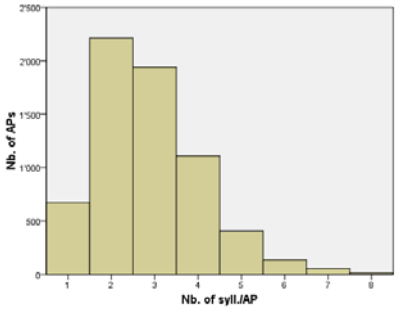

Figure 2. Number of APs according to their syllabic length.

3.1.3. *IP length*

In total, the corpus contains 3'895 IPs. The mean size of an IP is 4.79 syllables. As seen in Figure 3, most of the IPs that comprise our corpus are from 2 to 7 syllables. Beyond 10 syllables, the number of syll./IP decreases significantly:

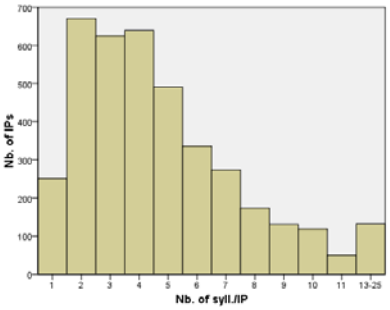

Figure 3. Number of IPs according to their syllabic length.

3.2. Statistical Analysis

We wanted to go further and provide a statistical analysis of the data presented in this paper. We report the results we obtained by testing the effects of age, articulation rate, order condition (logical vs. non-logical) and level of complexity (easy, medium, difficult) on accentuation (prominence rate) and phrasing (length of prosodic units). Three Generalized Estimated Equations (with repeated measures) models were run, for the first with the rate of prominence as a dependent variable, for the second with AP length as a dependent variable, and for the third with IP length as a dependent variable. Age, local articulation rate (mean syllabic duration excluding the silent pause for each AP and each IP), order condition and level of complexity were entered as predictors.

First, results indicate that none of the 4 predictors mentioned above had any effect on prominence rate. In other words, young speakers did not behave differently from the older ones, and both groups did not produce less pitch accents when they increased the pace at which they talked, or when they told an easy story in the logical order compared with a difficult story in the same order or not.

However, the level of complexity of the task had a significant effect on AP length ($p < 0.01$). Post hoc tests reveal that APs were shorter for level 3 stories than for level 1 stories ($p < 0.01$) but that they did not show any differences for level 2 stories. In addition, it appears that articulation rate had an effect on AP length: the faster the speaker articulated, the longer his/her APs ($p < 0.001$). We should note that there was an interaction between articulation rate and age ($p < 0.001$). This effect was not the same for young speakers as for older ones. As seen in Figure 4, older speakers had a longer syllabic duration when the prosodic constituent was short compared with the young speakers. Note that this difference in length tended to weaken as the length of the prosodic unit increased:

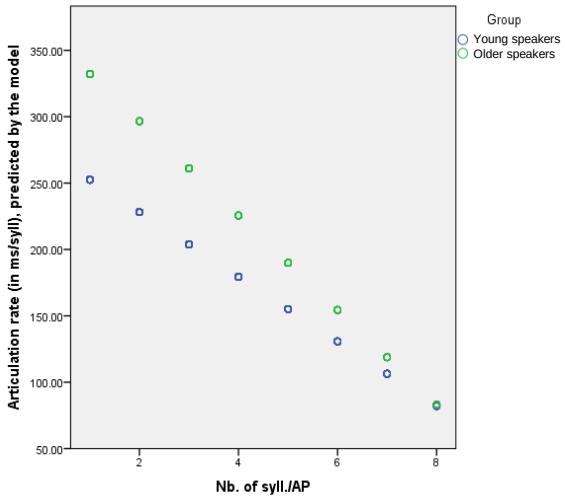

Figure 4. Number of syll./AP according to the age of the speakers and articulation rate, expressed in ms/syll., and predicted by the model.

Finally, the analysis revealed that IP length was not influenced by the condition or level of difficulty. However, it was influenced by articulation rate ($p < 0.001$), which interacted with age ($p < 0.001$). As was the case for the AP, long IPs presented a shorter mean duration than short IPs, which, on the other hand, presented a higher syllabic mean duration.

4. Discussion

The results presented in the preceding section are interesting with respect to our knowledge of French prosody. First,

regarding prominence and disfluency rate, the results found in this study can be compared with other corpus-based analyses of French. For example, out of the 21'161 syllabic intervals in C-PROM-Task, 35% were annotated as prominent while 11.6% were annotated as disfluent. Out of the 17'778 syllables in the C-PROM corpus [1], 4'570 syllables were annotated as prominent (26%) and 805 syllables (4.5%) were associated with a disfluency. In the Rhapsodie corpus [10], which contains 45'192 syllables, the prominence rate is 41% while the disfluency rate is slightly less than 8%.

Next, regarding phrasing, our results are in agreement with the ones obtained by [21], who found that the average length of the minor prosodic units that she calls "rhythmic groups" (and which correspond to the units we call APs) was 3.5 syllables in her corpus (2.85 syll./AP in our dataset). She also found that 80% of the APs in her data was composed of 2 to 4 syllables. Based on our data, we made exactly the same observation. The results obtained in our study are also in agreement with [6]'s description of French. Indeed, the authors found that an AP is composed of 3.5 to 3.9 syllables on average. Our data also confirm the idea that in French an AP cannot contain more than 8 syllables ([22]). Very little work is available for IP and it is therefore impossible to compare our results with other work.

Preliminary analyses examining the impact of some factors such as order condition, level of difficulty, articulation rate or age on accentuation and phrasing led to quite unexpected conclusions. On the one hand, it appeared that prominence rate was not influenced by one of the 4 factors, while AP and IP lengths varied according to articulation rate (the effect differing by gender). Additional analysis, not detailed here, revealed a significant effect of age on articulation rate: older speakers articulated much more slowly than younger speakers. These results are in agreement with previous studies on articulation rate in French [23].

5. Conclusions

The aim of this paper was to present C-PROM-Task, a new prosodically annotated dataset for the study of French prosody. The entire database is more than 2 hours long and contains speech by a group of young and a group of older native French-speakers from Switzerland. Recordings of controlled story-telling were first transcribed and aligned. They were then annotated for the study of prominent and disfluent syllables. Kappa measures were used to assess reliability between the annotators, which was judged to be fair to substantial. Recordings were also segmented in minor and major prosodic units, here called Accentual Phrases and Intonational Phrases. From a descriptive analysis of the data we were able to confirm previous findings. On average, one syllable carries a pitch accent every three or four syllables; 10% percent of a speaker's syllables is associated with a disfluency; AP length comprises between 3 and 4 syllables, and cannot exceed 8 syllables, while IP length is more sensitive to variation. Finally, the effects of many factors on articulation rate were tested, and it appeared that only age and constituent length had an effect on this prosodic variable.

6. Acknowledgments

This work was supported by the SNF under Grant No. 100012-140269, hosted at Neuchâtel University.

7. References

[1] Avanzi, M., Simon, A. C., Goldman, J.-P. and Auchlin, A. "C-PROM. An annotated corpus for French prominence studies", Proc. of Prosodic Prominence, Speech Prosody Satellite Workshop, 2010.

[2] Blanche-Benveniste, C., Bilger, M., Rouget, C. and van den Eynde, K. "Le français parlé. Etudes grammaticales", Paris, Éditions du CNRS, 1990.

[3] Morel, M.-A. and Danon-Boileau, L. "Grammaire de l'intonation : l'exemple du français", Paris/Gap, Ophrys, 1998.

[4] Groupe de Fribourg. "Grammaire de la période", Bern, Peter Lang, 2012.

[5] Ladd, R. "Intonational Phonology", Cambridge University Press, 1996.

[6] Jun, S. A., and Fougeron, C. "Realizations of Accentual Phrase in French intonation", Probus, 14: 147-172, 2002.

[7] Delais-Roussarie, E. et al. "Developing a ToBI system for French", in Frota, S & Prietro, P. (eds). Intonational Variation in Romance, Oxford University Press, in press.

[8] Lacheret, A. "La prosodie des circonstants", Paris/Leuven, Peeters, 2003.

[9] Delais-Roussarie et al. "Outils d'aide à l'annotation prosodique de corpus". Bulletin PFC, 6: 7-26, 2006.

[10] Lacheret A., Kahane, S. and Pietrandrea, P. "Rhapsodie: a Prosodic and Syntactic Treebank for Spoken French". New York/Amsterdam, Benjamins, in press.

[11] Champagne-Lavau, M., Fossard, M., Martel G., Chapdelaine, C., Blouin, G., Rodrigez, J-P. and Stip, E. "Do patients with schizophrenia attribute mental states in a referential communication task?", Cognitive Neuropsychiatry, 14(3): 217-239, 2009.

[12] Clark, H. H. and Wilkes-Gibbs, D. "Referring as a collaborative process", Cognition, 22, 1-39, 1986.

[13] Boersma, P. and Weenink, D. "Praat: doing phonetics by computer (Version 5.5)". www.praat.org, 2013.

[14] Goldman, J.-Ph. "EasyAlign: an automatic phonetic alignment tool under Praat", Proc. of Interspeech, 3233-3236, 2011.

[15] Avanzi, M., Obin, N., Lacheret-Dujour, A. and Victorri, B. "Toward a Continuous Modeling of French Prosodic Structure: Using Acoustic Features to Predict Prominence Location and Prominence Degree", Proc. of Interspeech, 2011.

[16] De Looze, C. and Hirst, D. "Detecting key and range for the automatic modelling and coding of intonation", Proc. of the Speech Prosody Conference, 135-138, 2008.

[17] Delais-Roussarie, E. and Post, B. "Unités prosodiques et grammaire de l'intonation : vers une nouvelle approche", Actes des 27èmes JEP, Avignon, 2008.

[18] Cohen, J. "A Coefficient of Agreement for Nominal Scales", Educational and Psychological Measurement, 20(1): 37-46, 1969.

[19] Fleiss, J. L. "Measuring Nominal Scale Agreement among Many Raters", Psychological Bulletin, 33: 613-619, 1973.

[20] Avanzi, M. "Note de recherche sur l'accentuation et le phrase du français à la lumière des corpus", Tranel,, 2013.

[21] Delais-Roussarie, E. "Pour une approche parallèle de la structure prosodique. Étude de l'organisation prosodique et rythmique de la phrase française". PhD thesis, Toulouse-le Mirail University, 1995.

[22] Martin, P. "Prosodic and Rhythmic Structures in French", Linguistics, 25: 925-949, 1987.

[23] Schwab, S. "Les variables temporelles dans la production et la perception de la parole", PhD thesis, Geneva University, 2007.