Naïve listeners' perception of prominence and boundary in French spontaneous speech

Guillaume Roux[1], *Roxane Bertrand*[1], *Alain Ghio*[1], *Corine Astésano*[2]

[1] Laboratoire Parole et Langage UMR7309-CNRS, Aix en Provence, France.
[2] URI Octogone-Lordat EA4156, Université Toulouse Jean Jaurès, Toulouse, France.

guillaume.roux@lpl-aix.fr, roxane.bertrand@lpl-aix.fr, alain.ghio@lpl-aix.fr, corine.astesano@univ-tlse2.fr,

Abstract

Our main goal here is to show how consistently, naïve listeners can identify different levels of prominences and boundaries in French spontaneous conversational data. We first present why and how the corpus investigated here, consisting in 133 utterances extracted from the Corpus of Interactional Data (CID), was created. 73 naïve listeners judged prominences and boundaries using three levels of prominence and boundary strength ("none", "weak" and "strong") during two separate real-time evaluation tasks. Prominence-Scores and Boundary-Scores reveal a good reliability between the listeners. After establishing a "gold" based on the strong agreement between two experts annotation, we briefly examine the extent to which naïve judgments are in line with expert annotations. The comparisons reveal a positive trend that encourages future investigations.

Index Terms: speech perception, prominences, boundaries, spontaneous speech, naïve listeners.

1. Introduction

This work focuses on how naïve listeners judge prominences and boundaries in French spontaneous conversational speech. Description of French phonological models [1 2, 3, 4] proposed a final accent (FA) and an optional Initial Accent (IA) [4, 5]. FA and IA are seen as right and left markers of the prosodic structures. This latter is defined by two levels of phrasing: a minor prosodic phrase or accentual phrase (AP) and an intonational phrase (IP), the highest in the hierarchy. Recent studies have experimentally shown a third level of phrasing (intermediate phrase-ip) [6]. Although ip requires to be refined, namely for spontaneous speech, [7] have suggested that a third level seemed necessary in spontaneous data. Plus, if very few studies provided evidence for differentiating ip from IP in French, Noun Phrases (NP) seem to be a valuable candidate for exhibiting an ip. In order to compare our results with those on the same corpus by [5] (containing NP), we only selected spontaneous extracts containing some NP patterns in our corpus.

The present study is a part of a larger project which aims at improving our comprehension of the relationship between accentuation and phrasing. Indeed, FA is postlexical and syncretic to boundaries in French. Numerous studies assume that prominences and boundaries can be seen as the same underlying phenomena [8, 9]. We present here the preliminary results on a perception study on a French spontaneous conversational speech data. The first question we address deals with the ability of naïve untrained French listeners to consistently perceive boundaries and prominences.

Listeners' prominence and boundary perception in spontaneous speech has been studied in Dutch [10, 11], in American English [12, 13, 14, 15, 16] and in French [17]. These works showed that listeners judge efficiently boundaries and prominences in spontaneous speech with a strong agreement between listeners, higher agreement with boundaries than with prominences, and also strong agreement throughout naïve listeners and trained listeners [18, 19, 15]. Correlations have also been established with phonetic cues such as duration and intensity [16], pre-nuclear/nuclear prominences [15] or between spontaneous and data-driven speech in French, on radiophonic data and Maptask [17].

This study aims at taking into account to if naïve untrained listeners are consistent in perceiving prominences and boundaries in French spontaneous conversational data. We hypothesize that listeners perceive boundaries with a higher agreement than prominences as previous studies have shown [11, 14 among others]. Also, most of previous studies have been made on boundary and prominence perception according to only two levels: presence or absence. In this work, we added a third level for both boundary and prominence tasks. Listeners were asked to identify them on three levels as: none, weak and strong. By using a more fine-grained level of boundaries and prominences, we also examine how consistently each level is perceived. We hypothesize that strong boundaries and strong prominences will yield the best agreement. We then test agreement on prominences and boundaries between naïve listeners and we show first results comparing these perceptual results with experts' annotation.

2. Method and analysis

2.1. Corpus

133 utterances were extracted from the CID (Corpus of Interactional Data) [18] including 2778 tokens and 3395 syllables. The selection had short and long utterances (duration between 3 and 15 seconds) from the 16 speakers of the CID.

The CID is an audio video recording of French spontaneous face-to-face conversations. Speakers were recorded in an anechoic room and each of them was equipped with a headset microphone enabling the recording of the two speakers' voice on two different sound tracks. This results in a high quality of speech allowing a very fine-grained analysis at the different levels of speech. From an orthographic transcription, numerous annotations were performed at the different linguistic levels [19, 20]. For this study, we used the

morphosyntactic annotation provided by the stochastic parser Marsatag [21] which provides for each part-of-speech token an automatic annotation of its morpho-syntactic category. From this annotation, we extracted all the NP in the CID (10735) and then selected the three most frequent patterns (Table 1).

Table 1: *The three most frequent complex morphosyntactic patterns firstly selected.*

Morphosyntactic pattern	D_Adj_N	D_N_Adj	D_N_Spd_N
Frequency	462	403	430

Then syntactic functions of NP have been manually annotated by two experts. A Cohen's kappa has been calculated to estimate the reliability between their judgments. The kappa score was 0.877 reflecting a strong agreement. Direct object was the most frequent function (51.5% of the obligatory functions), with 514 NP coded as a direct object. In those 514 direct objects, we selected for this present study only the ones in utterances easily audible, understandable, with no syntactic and semantic ambiguity and with the less disfluencies as possible.

2.2. Experiment perception procedure

73 naïve listeners without knowledge in phonetics and phonology have been recruited. They are between 18 and 55 years old and do not have any auditory problems. 8 listeners could simultaneously take part in the experiment thanks to the number of computers available in the computer classroom.

The experiment is led on PERCEVAL software [22], a computer-driven system for experimentation on auditory and visual perception developed by the team LPL-CNRS of Aix-en-Provence, France. The auditory part is led on a program called LANCELOT, depending on PERCEVAL. Two scripts have been created, one for each task. A three degree evaluation has been prepared, "rien, faible, fort" (none, weak, strong). "None" is selected by default.

The prominence task is based on syllables and not chunks [11, 15 among others]. The utterances were cut out in syllables separated by a space, orthographically transcribed, without any punctuation nor capital letters. The points of judgment were put under each syllable (Figure 1). For the boundary task, the utterances were cut out in words separated by a space, without any punctuation nor capital letters. The points of judgment were put between each word (Figure 2). For each task, except inside the NP, disfluencies such as repetition, filler pause, syllabic lengthening can occur in the utterances.

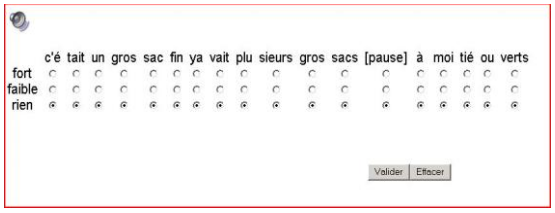

Figure 1: *Example of an utterance from the prominence task.*

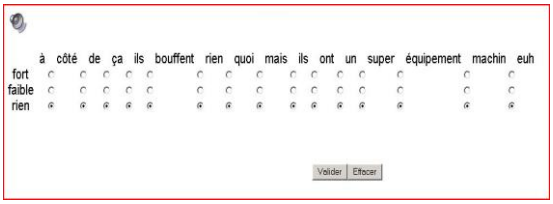

Figure 2: *Example of an utterance from the boundary task.*

Listeners were divided in 4 blocks of 18, each block divided in 2 groups of 9 participants. Groups of 9 were constituted to balance the number of listeners in each group and to have a pair number in each block to better establish boundary and prominence scores. The listeners were assigned to two tasks (prominences and boundaries) on the same utterances (random order in each of the task), but there were 4 different set of 33 utterances, one set per block. Before the experiment, they were given orally some instructions by a supervisor, and the instructions related were detailed on the screen before each task. For the prominence task, they were asked to focus on the musical salience of syllables [16]. For the boundary task, they were asked to focus on a feeling of break in the utterance. They were provided for two hours to do the two tasks. After the first task is done, the listener had to call the supervisor to put on the second task. They could have a break between each task and in the middle of each task. Each listener could listen to the sound of the utterance a maximum of 10 times, clicking on the sound logo. Then, they had to click on the buttons of salient syllables or breaks according to the task they were doing. After their judgment is done, they clicked on "Valider" (Valid) to valid their evaluation or they could click on "Effacer" (Clear) to start clear their selection.

In addition to the naïve transcription [15], an expert transcription for 43 utterances, resulting in their agreement, has been provided to compare with the different levels of boundaries and AI AF based on theoretical models [1, 2, 3, 4]. The set of 43 utterances is extracted from the utterances judged by the naïve listeners.

3. Results

The reliability of the boundaries and prominences perception by 73 naïve listeners is measured by a score agreement for each syllable and each tokens' interval according to the strength level (none, weak, strong). As illustrated in figures 3 and 4, a Prominence-Score (P-Score) and a Boundary-Score (B-Score) has been calculated such as in [13]. A score of 0 is when "none" is chosen, a score of 1 is when "weak" is chosen and a score of 100 is when "strong" is chosen. Yet, on figures 3 and 4, the scale is the same (from 0 to 10) for a better view.

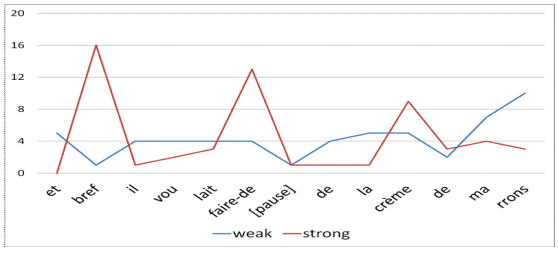

Figure 3: *P-Scores on utterance n°3 on the 4th group.*

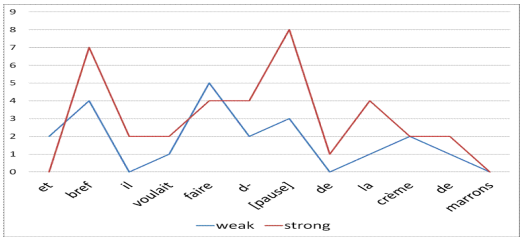

Figure 4: *B-Scores on utterance n°3 on the 4th group.*

B-Scores showed that on all judged tokens' intervals, ones judged as "none" constitute the most important category (Figure 5). 84% of the tokens' intervals are judged as "none". 12% of tokens' intervals were judged with a boundary. 4% of the intervals are ambiguous and judged as only 50% weak or strong or none. In the prominence task, 70% of the syllables are judged as "none" and is the most important category (Figure 6). 24% of syllables are judged with the presence of prominences, weak and strong mixed. 6% of the judgments are ambiguous and judged as only 50% weak or strong or none.

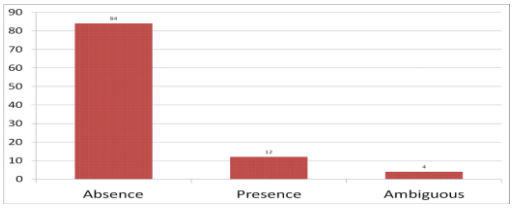

Figure 5: *Frequency of B-Scores for presence or absence and of bad judged tokens' intervals.*

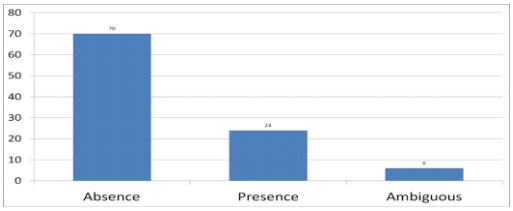

Figure 6: *Frequency of P-Scores for presence or absence and of bad judged syllables.*

At the three levels of judgment, figures 7 and 8 show 6% of strong prominences identified and 1.3% of strong boundaries. Prominences are more likely to be judged as strong than boundaries. Weak prominences and weak boundaries are the less identified. 0.6% of the syllables have weak prominences and 0.9% of the tokens' intervals are weak boundaries.

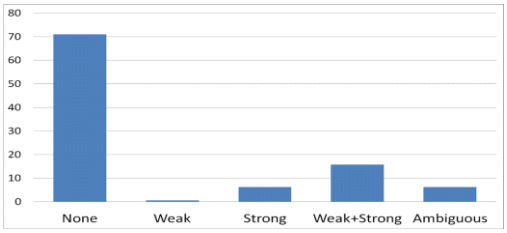

Figure 7: *Three-leveled frequency of sites of prominences well judged and ambiguous.*

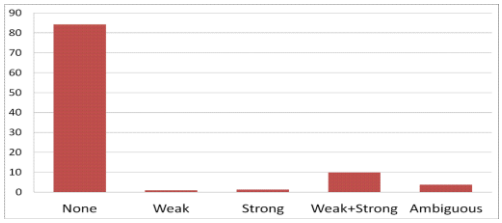

Figure 8: *Three-leveled frequency of sites of boundaries well judged and ambiguous.*

Weak prominences and boundaries mostly judged are the less number of prominences and boundaries. 1% of ambiguous half judged syllables concerns weak prominences against 0.6% well judged weak prominences. 0.9% of ambiguous half judged tokens' intervals concern a weak boundary against 0.9% well judged weak boundaries. Weak prominences and boundaries are the less well judged and weak prominences are less well identified than weak boundaries. Moreover, 2% of ambiguous syllables concern strong prominences against 6.3% well judged ones. 0.96% of ambiguous tokens' intervals concern strong boundaries against 1.3% well judged ones.

As we said above, an expert annotation has been fulfilled. The inter-annotator agreement between the 2 raters was evaluated using a Cohen's kappa. The score is 0.85 for IP/ip (z=29.5), 0.85 (z=29.5) for AI/AF and 0.91 (z=29.7) for AP. As for naïve experiment, a score has been established between the two experts in order to make the comparison easier (AI=100, AF=1, ip=1, IP=100 and AP=1, none=0). In a first attempt to show whether naïve perception and expert annotation is in line, a Spearman's correlation test has been calculated. The correlation between naïves' judgments on boundaries and experts IP/ip is with r= 0.40; between naïves' judgments on boundaries and experts AP is r=0.33; between naïves' judgments on prominences and AI/AF it is r=0.44. Correlations are quite mixed. Results from X^2 also show that the frequency of prominences and boundaries identified by naïve listeners and experts are different. $X^2 = 31.362$, df = 1, p-value <0.001 with 35% of prominences for experts and 24% for naïve listeners. $X^2 = 23.5104$, df = 1, p-value<0.001 with 18% of boundaries for experts and 10% for naïve listeners.

4. Discussion

This study was designed to observe the consistency of naïve listeners' judgment of French spontaneous speech on prominences and boundaries in a difficult experiment based on two tasks using three levels of judgment scores. In the first step, we established P-scores and B-scores enabling us to measure a good agreement between naïve listeners in French spontaneous speech in spite of the difficulty of the tasks. Scores show a strong agreement on prominences and boundaries judged as "none" (absence). Prominences and boundaries are also well judged since we only found 6% of ambiguous judgments for the prominence task and 4% for the boundary task. However, it has been observed that with three judgment levels, the distinction of mostly scored weak prominences and boundaries and mostly scored strong ones is less easy. There are more ambiguous judgments than good ones for weak prominences and boundaries. Strong prominences mostly judged are most numerous than weak ones. Strong prominences and boundaries mostly judged seem more consistent in judgment than weak ones. But a task using three levels of judgment seems to be relatively difficult for

naïve listeners although when only presence and absence are judged, scores a really good, confirming previous studies [11, 14 among others]. Naïve listeners are able to consistently judge prominences and boundaries in spontaneous speech.
The second part of our work is based on the comparison with listeners' judgments and the annotation of two experts on a pool of 43 utterances. They annotated Initial Accent (IA), Final Accent (FA), Accentual Phrases (AP), intermediate phrases (ip) and Intonational Phrase (IP) [1, 2, 3, 4]. Cohen's kappa between the two raters has shown very good agreement. The correlation between the two experts and the naïve speakers indicates that the judgments are quite mixed.

Yet, our results show consistent agreement within naïve listeners' judgments and within in experts' judgments. The question then arises as to why the correlation between naïves and experts is unclear/weak? We hypothesize that both populations do not judge comparable events. It is necessary to qualitatively evaluate what influences naïve listeners' judgment. For example, disfluences could play a major role in the way naïve listeners judge some prosodic phenomena [12], which is not always the case of experts. Two ways of future investigation are then planned: the first concerns Part-of Speech and boundary/prominence relationship. The second is to focus on naïve listeners' and experts annotators within NP.

5. Conclusion

This study allowed us to confirm that French naïve listeners can perceive prominences and boundaries with a consistent agreement in a difficult three-leveled judgment score over two tasks. It also reveals that, according to what has been observed on other languages, boundaries have less ambiguous judgments made by naïve listeners than prominences, but that strong prominences are most numerous than weak ones. The link between the experiment on naïve listeners and the experts' annotation shows a positive trend that needs to be more investigated.

6. Acknowledgments

This study is supported by the Agence Nationale de la Recherche grant ANR-12-BSH2-0001 (PI: Corine Astésano). We want to thank Stéphane Rauzy, Carine André, Laury Garnier, our listeners and the speakers of the CID corpus.

7. References

[1] S. A. Jun and C. Fougeron, "Realizations of accentual phrase in Frenc*h intonation," Probus no 14(1), pp. 147-172*, 2002.

[2] B. Post, "Tonal and phrasal structures in French intonation," *Thesis*, The Hague, 2000.

[3] A. Di Cristo, "La prosodie au carrefour de la phonétique, de la phonologie et de l'articulation formes-fonctions," *Travaux interdisciplinaires du Laboratoire parole et langage d'Aix-en-Provence*, vol. 23, pp. 67-211, 2004.

[4] P. Welby, "French intonational structure: Evidence from tonal alignment," *Journal of Phonetics*, vol. 34, pp. 343-371, 2006.

[5] C. Astésano, E. Bard and A. Turk, "Structural influences on Initial Accent placement in French," *Language and Speech*, no 50(3), pp. 423-446, 2007.

[6] A. Michelas and M. D'Imperio, "Uncovering the role of the intermediate phrase in the syntactic parsing of French," in *17th International Congress of Phonetic Sciences*, pp. 1374-1377, 2011.

[7] I. Nesterenko, S. Rauzy and R. Bertrand, "Prosody in a corpus of French spontaneous speech: perception, annotation and prosody syntax interaction", in *Speech Prosody 2010-Fifth International Conference*, Chicago, pp. 1-4, 2010.

[8] J. Vaissière, "Rhythm, accentuation and final lengthening in French," in J. Sundberg, L. Nord & R. Carlson (Eds.), *Music, language, speech and brain*, New York, MacMillan Press, pp. 108-120, 1991.

[9] M. E. Beckman, "Evidence for Speech Rhythms across Languages," in Tohkura et al. (Eds.), *Speech Perception, Production and Linguistic Structure*, Tokyo, pp. 457-463, 1992.

[10] B. M. Streefkert, L. C. W. Pols and L. F. M. ten Bosch, "Prominence in read aloud sentences as marked by listeners and classified automatically," *Proceedings of IFA* (Amsterdam), pp. 101-116, 1997.

[11] J. Buhmann, J. Casper, V. J. Heuven van, H. Hoekstra, J. P. Martens and M. Swerts, "Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus," in *Proceedings of LREC 2002* (Las Palmas), pp. 779-785, 2002.

[12] J. Cole, Y. Mo and S. Baek, "The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech," *Language and Cognitive Processes*, no 25(7), pp. 1141-1177, 2010.

[13] T-J. Yoon, S. Chavarria, J. Cole and M. Hasegawa-Johnson, "Intertranscriber Reliability of Prosodic Labeling on Telephone Conversation using ToBI," *Proceedings of Interspeech 2004* (Jeju), pp. 2722-2732, 2004.

[14] J. Cole, T. Mahrt and J. I. Hualde, "Listening for sound, listening for meaning: Task effects on prosodic transcription," *Speech Prosody* 7, Dublin, Ireland, 2014.

[15] Y. Mo, J. Cole and E.-K. Lee, "Naïve listeners' prominence and boundary perception," *Speech Prosody 2008*, Campinas, Brazil, pp. 735-738, 2008.

[16] Y. Mo, "Duration and Intensity as perceptual cues for naïve listeners' prominence and boundary perception," *Speech Prosody 2008*, Campinas, Brazil, pp. 739-742, 2008.

[17] C. L. Smith, « Naïve listeners' perceptions of French prosody compared to the predictions of theoretical models," in H-Y. Yoo & E. Delais-Roussarie (Eds.), *Proceedings from IDP 2009* (Paris), pp. 335-349, 2009.

[18] R. Bertrand, P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valdere and S. Rauzy, "Le CID – Corpus of Interactional Data-: protocols, conventions, annotations", *Travaux interdisciplinaires du Laboratoire parole et langage d'Aix-en-Provence (TIPA)*, vol. 25, pp. 25-55, 2006.

[19] P. Blache, R. Bertrand and G. Ferré, "Creating and exploiting multimodal annotated corpora: the toma project," in *Multimodal Corpora*, pp. 38-53, 2009.

[20] P. Blache, R. Bertrand, B. Biggi, E. Bruno, E. Cela, R. Espesser, G. Ferré, M. Guardiola, D. Hirst, R. Muriasco, J. C. Martin, C. Meunier, M. A. Morel, I. Nesterenko, P. Nocera, B. Palaud, L. Prévot, B. Priego-Valverde, J. Seinturier, N. Tan, M. Tellier and S. Rauzy, "Multimodal annotation of conversational data," in *Proceedings of Linguistic Annotation Workshop*, Uppsala, Sweden, pp. 186-191, 2010.

[21] S. Rauzy, G. Montcheuil and P. Blache, "Marsatag, a tagger for French written texts and speech transcriptions," in *Proceedings of Second Asian Pacific Corpus linguistics Conference*, Hong Kong, Hong Kong, pp.220, 2014.

[22] C. André, A. Ghio, C. Cavé and B. Teston, "PERCEVAL: a Computer-Driven System for Experimentation on Auditory and Visual Perception," *Proceedings of XVth ICPhS*, Barcelona, Spain, pp. 1421-1424, 2003.