INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

How does the absence of shared knowledge between interlocutors affect the production of French prosodic forms?

Amandine Michelas[1], *Cecile Cau*[1], *Maud Champagne-Lavau*[1]

[1]Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France

michelas@lpl-aix.fr, cecile.cau@etu.univ-amu.fr, maud.champagne-lavau@univ-amu.fr

Abstract

We examine the hypothesis that modelling the addressee in spoken interaction affects the production of prosodic forms by the speaker. This question was tested in an interactive paradigm that enabled us to measure prosodic variations at two levels: the global/acoustic level and the phonological one. We used a semi-spontaneous task in which French speakers gave instructions to addressees about where to place a cross between different objects (e.g., *Tu mets la croix entre la souris bordeau et la maison bordeau*; 'You put the cross between the red mouse and the red house'). Each trial was composed of two noun-adjective fragments and the target was the second fragment. We manipulated (i) whether the two interlocutors shared or didn't share the same objects and (ii) the informational status of targets to obtain variations in abstract prosodic phrasing. We found that the absence of shared knowledge between interlocutors affected the speaker's production of prosodic forms at the global/acoustic level (i.e., pitch range and speech rate) but not at the phonological one (i.e., prosodic phrasing). These results are consistent with a mechanism in which global prosodic variations are influenced by audience design because they reflect the way that speakers help addressees to understand speech.

Index Terms: language production, audience design, shared knowledge, prosody, French

1. Introduction

A controversial issue in research on language production is to what extent speakers serve the needs of their addressee when choosing particular linguistic forms (e.g., referential expressions, prosodic forms) in spoken interaction. Two theoretical positions can be defined. According to the first approach (the speaker internal view), most of linguistic choices are motivated by the speaker's own experience and rely primarily on his/her private knowledge [1, 2]. Within this framework speakers mainly consider referential expressions, for example, from their own perspective even when they known that these expressions are inaccessible to their addressee [1]. By contrast, the second approach assumes that speakers formulate utterances by consulting information that is mutually shared with their addressee (e.g., they choose referring expressions such as pronouns on the basis of shared knowledge with the addressee [3, 4, 5, 6]). The view that language production is mainly addressee-oriented and not speaker-oriented is often referred to the *audience design hypothesis* in the literature.

However, in a general manner neither of these two conflicting views is tenable. Strictly speaker-internal view cannot explain that at least in some circumstances speakers select referring expressions on the basis of the knowledge, intentions or goals of their addressee [5, 6] and a strictly audience design view is difficult to defend since maintaining an incrementally updated model of what the listener knows is a cognitively demanding task [7].

Interestingly, prosodic variations, which refer to melodic and rhythmic variations, can be analysed within these two conflicting frameworks. According to the speaker-internal view, prosodic variations reflect situations in which speaker's production is facilitated (e.g., the duration of a word that has been already mentioned is reduced because it is easier to mentioned the word again [8, 9, 10]) while according to the audience design hypothesis prosodic variation is addressee-given because it reflects the way that the speaker helps his/her addressee to understand speech (e.g., when a word is particularly relevant in context, the speaker select prosodic forms that make it salient for the addressee [11]). One important aspect when considering prosodic variation is that these variations can be analysed at two linguistic levels: the acoustic level and the phonological level of speech. The acoustic level of prosody concerns variations of acoustic/phonetic cues such as duration of segments or fundamental frequency. These phonetic cues assume two main functions: they encode abstract prosodic forms (i.e., prosodic constituents are manly encoded by variations of duration and specific melodic contours in French) and give rise to global prosodic variations such as pitch range and speech rate that can vary according to speaker-specific or situational factors. The phonological level of prosody concerns both the prosodic phrasing (i.e., the grouping of words into abstract prosodic constituents) and the positioning of accentual prominence within sentences. It serves important communicative functions at the grammatical and pragmatic levels.

Among its pragmatic functions, abstract prosodic forms encode informational status of words. For instance, in French, new contrastive information tends to be phrased in a separate abstract prosodic unit (the accentual phrase or AP) while already mentioned information tends to be produced together with adjacent words [12,13,14]. In an interactive paradigm in which a director had to indicate noun-adjective pairs of items to an addressee, Michelas et al. [14] showed that participants parsed the target noun in the same AP when it was identical to the noun in the preceding fragment (e.g., *BONBONS marron* followed by [*BONBONS violets*]$_{AP}$) while they parsed it in a separate AP when it contrasted with it to warn their interlocutor that this noun constituted a contrastive entity (e.g., *BOUGIES violettes* followed by [*BONBONS*]$_{AP}$[*violets*]$_{AP}$).

Although such pragmatic function of prosody is well known, the mechanism behind it is not fully understood. An unresolved question is whether audience design affects these prosodic variations or not. Thus, the present research was designed to examine how modelling the addressee affects the production of prosodic forms. We had two goals. First, to

http://dx.doi.org/10.21437/Interspeech.2017-1430

determine whether the manipulation of absence/presence of shared knowledge have an effect on prosodic variations of words and second to determine which level of prosodic forms is affected by the knowledge between interlocutors (the phonological level of prosody vs. the global/acoustic one).

2. Method

2.1. Participants

Twenty two participants (16 women) between 17 and 47 years old (mean = 20.92) participated in the experiment. All reported having no neurological or hearing impairment.

2.2. Materials

Participants performed an interactive task involving cooperation between a director (the participant) and an addressee (the experimenter). The director had to give instructions to the addressee about where to place a cross between different objects (e.g., *Tu mets la croix entre la souris bordeau et la maison bordeau*; 'You put the cross between the red mouse and the red house'). Each trial was composed of two noun-adjective fragments (e.g., red mouse vs. red house). The target was the second fragment (e.g., red house).

The director (the participant) had to indicate 78 critical pairs of noun-adjective fragments to the addressee. The objects were chosen so that we controlled for the words produced by the participants (e.g. the number of syllables, the syllable structure and the phoneme characteristics of words).

We developed two experimental manipulations. First, to control for the presence vs. absence of shared knowledge between participants, we manipulated the speaker's knowledge about whether his/her addressee shared or didn't share the same objects as him/her. To do so, in the ***not-shared knowledge condition***, target items appeared in black boxes and participants knew that they could have different shape/colour for objects in those boxes while in the ***shared knowledge condition***, items appeared in white boxes and participants knew that it meant that their interlocutor shared the same objects as him/her. Second, in order to induce phonological prosodic variations, we manipulated the informational status of the second fragment since contrastive status of words is well known to lead to differences in AP phrasing [12, 13, 14]. To do so, we manipulated whether the noun or the adjective in the 2nd fragment was the same or contrasted with the noun or the adjective in the 1st fragment. We obtained three types of informational status: (1) the noun in the 2nd fragment contrasted to the noun in the 1st fragment while the adjective was keep constant (e.g., *souris bordeau* vs. *maison bordeau*; ***noun contrast condition***), (2) the adjective in the 2nd fragment contrasted to the adjective in the 1st fragment while the noun was keep constant (e.g., *maison violette* vs. *maison bordeau;* ***adjective contrast condition***, (3) both the noun in the 2nd fragment contrasted to the noun in the 1st fragment and the adjective in the 2nd fragment contrasted to the adjective in the 1st fragment (e.g., *souris violette* vs. *maison bordeau*; ***all contrast condition***). The two experimental manipulations leading to 6 conditions are illustrated in Table 1 for the target fragment *souris bordeau*. Figure 1 shows director's and addressee's views for the target item *maison bordeau* in the noun-contrast not-shared knowledge condition.

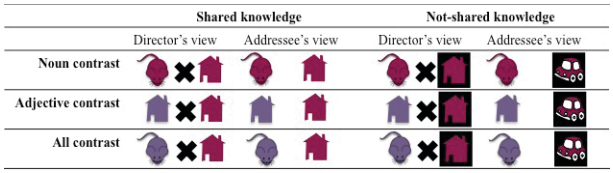

Table 1. *Objects corresponding to the maison bordeau 'red house' target fragment in the 6 experimental conditions.*

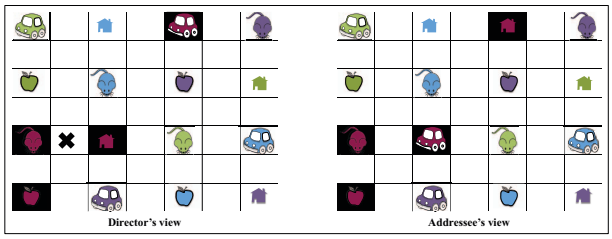

Figure 1. *The director and addressee's views for the maison bordeau 'red house' fragment in the noun-contrast not-shared knowledge condition. The director's view includes the cross (left panel) while the addressee's view doesn't (right panel).*

Each screen view was composed of 16 objects corresponding to the combination between 4 shapes and 4 colours. For each grid, two filler colours and two filler shapes were added to the shape and colour of critical pairs. In half of the screen views, addressees had to place the cross horizontally whereas in the other half, they had to place it vertically.

2.3. Procedure

Directors and addressees were seated, facing each other, both with a computer screen. They could not see each other's computer screen. Directors were asked to give instructions to addressees about where to place the cross by indicating to the addressee the shape and the colour of objects (e.g., *Tu mets la croix entre la souris bordeau et la maison bordeau*; 'You put the cross between the red mouse and the red house'). In the not-shared knowledge condition, directors were asked to help addressees to place the cross at the right place which corresponds to locations that directors had. Each screen view appeared simultaneously on the director and the addressee computer's screen. Once the addressee had indicated on his/her screen where to place the cross using the mouse, both participants were asked to simultaneously click on the space bar to make the next screen view appeared. Directors and addressees were recorded in a quiet room using a Zoom H4N Handy Recorder and a Headset Cardioid Condenser Microphone (AKG C520). We analysed directors' speech only.

To ensure that directors could identify and use a consistent label for each target object and colour, participants performed a familiarization phase during which they had to name what they saw described in pictures before the experimental phase. Participants were asked to remember these labels, as they would see the same pictures in the experimental phase. They began the experimental phase with a block of 6 practice trials.

2.4. Measures and annotations

The pairs containing neither disfluencies/hesitations nor object appellation errors were analysed for a total of 1461 pairs (85.1% of the original data). In these pairs, we analysed both

phonological (i.e., phrasing in APs) and global/acoustic prosodic forms (i.e., pitch range and speech rate).

We first analysed whether speakers produced the noun in the 2nd fragment in the same AP as the following adjective (e.g., [*maison bordeau*]$_{AP}$ 'red house') or in a separate AP (e.g., [*maison*] $_{AP}$ [*bordeau*]$_{AP}$). Pre-boundary lengthening and the presence of a typical fundamental frequency (f0) rise aligned with the last syllable of the AP are well known as the two main correlates of AP right boundaries [15, 16, 17, 18]. For this reason, we considered that an AP right boundary was actually produced by the speaker after the target noun if (i) the f0 maximum hertz value of the last syllable of the target noun was at least 10% higher than that of the preceding low inflection point in the f0 curve and if (ii) the duration of the last syllable of the target noun was at least 10% longer than that the preceding syllable (see [14] for this procedure).

We then analysed pitch range and speech rate as glocal/acoustic prosodic forms. Pitch range was observed in terms of pitch span variations that provide information on the excursion of melodic movements [19]. It was defined as the difference between maximum and minimum f0 values for the whole pair of noun-adjective fragments. Speech rate was measured per speaker by calculating the number of syllables per second on the whole pair of noun-adjective fragments.

In order to perform the prosodic measures described above, the following procedure was followed. All syllables of the two fragments were manually segmented and tagged as illustrated in Figure 2. We also manually tagged the low inflection point in the f0 curve near the beginning of the first syllable of the noun in the 2nd fragment (L in Figure 2) and the f0 maximum hertz value of the last syllable of the same target noun (H in Figure 2). We then automatically extracted all these values thanks to Praat scripts at the same time as f0 minima and maxima for the whole pair of fragments.

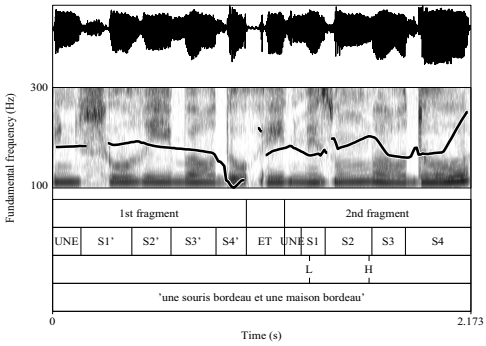

Figure 2. *Schema annotation for the pair une souris bordeau et une maison bordeau 'the red house and the red house'.*

3. Results

To test the statistical relevance of our two factors (type of knowledge and informational status), we employed mixed effects regression modelling. We used lme4 [20] implemented in R (R Development Core Team, 2017). For each model, the type of knowledge (shared knowledge vs. non-shared knowledge) and the informational status of target fragment (noun contrast, adjective contrast, all contrast) were entered into models as fixed factors. Following [21] the maximal random structure that allowed models to converge was observed.

3.1. Phonological phrasing in APs

The percentage of AP prosodic phrasing depending on whether the target noun of the target fragment was parsed as a separate AP (2APs) or within the same AP as the adjective (1AP) is shown in Figure 3.

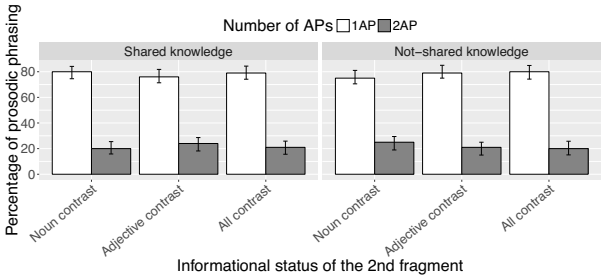

Figure 3. *Percentage of prosodic phrasing produced by participants. Error bars show a 95% confidence interval.*

As shown in Figure 3, there was a large bias in favour of the 1AP phrasing whatever the type of knowledge and the informational status of 2nd fragments. We ran a mixed effect logistic regression that included the phonological phrasing in APs produced by participants (1AP = 1, 2APs = 0) as the dependent variable. The model was structured as follows: APs~InformationalStatus*TypeOfKnowledge+(1|Participant)+(1|Item). Neither the effect of the two factors (effect of InformationalStatus: Chisq=0.19, p>.20; effect of TypeOfKnowledge: Chisq= 0.01, p>.20) nor the InformationalStatus x TypeOfKnowledge interaction (Chisq=1.96 ; p>.20) were significant.

3.2. Global prosodic variations

Mean of pitch span (in Hz) and mean of speech rate (in syl/s) are shown in Figure 4 and 5 respectively.

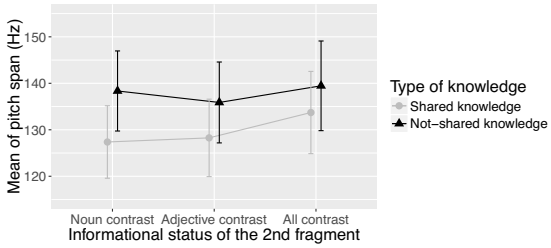

Figure 4. *Mean of pitch span depending on the type of knowledge and the informational status of the target fragment. Error bars show a 95% confidence interval.*

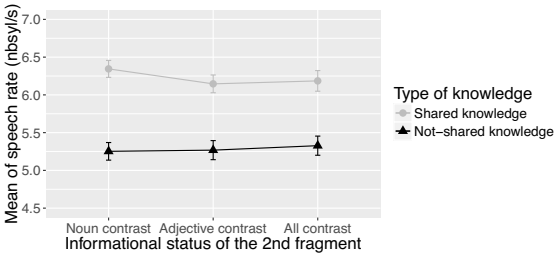

Figure 5. *Mean of speech rate depending on the type of knowledge and the informational status of the target fragment. Error bars show a 95% confidence interval.*

We ran two linear mixed models on logarithms of pitch span and on logarithms of speech rate. F0 values were log-transformed, in order to normalise the range of variability found both within and across speakers. Note that this scale is mathematically equivalent to a semitone conversion, since both scales are logarithmic transformations. Taking the log transformation also ensures a normal distribution of the residuals [22].

Pitch span analysis: The structure of the model was as follows:log_pitchspan~InformationalStatus*TypeOfKnowledge+(0+InformationalStatus|Participant)+(0+TypeOfKnowledge|Participant)+(1|Item). Pitch span of participants was smaller in the shared knowledge condition compared to the not-shared knowledge condition (when the intercept represents the noun contrast not-shared condition: t-value=-2.81). There was a main effect of TypeOfKnowledge (Chisq=4.24, p<.05) due to larger values in the not-shared knowledge condition compared to the shared knowledge one. The effect of InformationalStatus (Chisq= 3.29; p=.19) and the InformationalStatus x TypeOfKnowledge interaction (Chisq=3.67; p=.16) were not significant.

Speech rate analysis: The structure of the model was as follows:log_speechrate~InformationalStatus*TypeOfKnowledge+(0+InformationalStatus|Participant)+(0+TypeOfKnowledge|Participant)+0+InformationalStatus|Item)+(0+TypeOfKnowledge|Item). Speech rate was faster in the shared knowledge condition compared to the not-shared knowledge condition (when the intercept represents the noun contrast not-shared condition: t-value=9.61). There was a significant InformationalStatus x TypeOfKnowledge interaction (Chisq= 8.04, p<.05) due to faster speech rate in the noun contrast condition compared to the two other conditions only when the knowledge is not-shared between interlocutors.

4. Discussion

The goal of this study was to determine whether the presence vs. absence of shared knowledge between interlocutors affect the production of prosodic forms (both at the phonological and global/acoustic level of speech). Crucially, we found that the type of knowledge affected the global/acoustic prosodic features of participants' speech (i.e., pitch range and speech rate) but not the phonological prosodic ones (i.e., prosodic phrasing). Specifically when participants knew that their interlocutors did not share the same objects as them, they spoke slower and with larger pitch excursions compared to when they knew that their interlocutor shared the same objects as them. These findings establish, to our knowledge, a previously undocumented link between global/acoustic prosodic variations and audience design by showing that the absence of mutual knowledge between participants modified the global/acoustic prosodic forms produced by the speaker.

Considering the phonological level of speech, we did not observe an effect of the type of knowledge. However, the informational status of words did not induced AP phrasing modifications as we expected it. In fact our participants always parsed the noun of the 2nd fragment together with the adjective (i.e., they produced only one AP) whatever the informational status of words. This is quite surprising in light of previous studies showing that French speakers use AP phrasing to encode contrastive status of words within paradigms in which directors shared the same set of alternatives as their addressee [12,13,14]. Different from these studies, in the present experiment we manipulated the presence vs. absence of shared knowledge in addition to the informational status of words. It is possible that this additional manipulation has increased the cognitive load of speakers during the task and could have thus affected the way they used AP phrasing to encode informational status of words. Indeed, according to this manipulation, speakers had to constantly ask themselves whether their addressee had the same knowledge about the objects as them. This hypothesis is in line with a recent study [23] showing that when speakers are in a condition of double tasks, which increases their cognitive load, the accessibility of referents is decreased in their mental model. It is thus possible that our speakers didn't encode informational status of words whatever the type of knowledge they shared with their addressee because of the cognitively demanding task we used. Based on this assumption, two scenarios are possible regarding the effect of cognitive load on prosodic variations. First it is possible that an increase in cognitive load may result in difficulties for the speaker to take the addressee's perspective. Second, it is possible that an increase in cognitive load may affect the speaker's own discourse model by decreasing the accessibility of referents. This could explain why we did not observe an effect on prosodic phrasing while [12,13,14] did. Hence, future researches will have to determine whether cognitive load affect phonological prosodic variations in a more speaker-oriented or addressee-oriented way. Moreover French intonational phonology has given evidence for additional markers of the informational status of words, such as the presence of an initial f0 rise near the beginning of the AP [24, 25]. Thus, it could also be interesting for further work to investigate to what extent other French phonological prosodic forms than AP phrasing could be affected by an increase in cognitive load.

But whatever the effect of an increase in cognitive load on phonological prosodic forms, the absence of shared knowledge between interlocutors had for consequence to modify their global/acoustic prosodic features. These findings are consistent with a mechanism in which global prosodic variations are influenced by audience design because they reflect the way that speakers help addressees to understand speech. Our result can also be explained in the framework of Bard et al.'s Dual Process hypothesis [26], which had the advantage to reconcile purely speaker-oriented, and purely addressee-oriented accounts. This model suggests that two types of cognitive processes would be involved in interaction with addressees, one automatic and rapid, with no cognitive cost that refers to the speaker's own recent experience and the other slow and more cognitively demanding that involves a construction of the mental model of the addressee. Within this framework the decrease in speech rate and the increase of pitch range we observed would result from a slow and demanding cognitive process that would involve the construction of a mental model of the addressee. Thus, research must be further develop to determine whether a less demanding experimental procedure involving only one task (e.g., a procedure in which only noun contrast informational status would be tested) would also induce global/acoustic prosodic variations.

5. References

[1] B. Keysar, D. J. Barr, J. A. Balin and J. S. Brauner. "Taking perspective in conversation: The role of mutual

knowledge in comprehension". *Psychological Science*, *11*(1), 32-38, 2000.

[2] D. J. Barr and B. Keysar. "Perspective taking and the coordination of meaning in language use". In M. J.Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (pp. 901–938). Amsterdam: Elsevier, 2006.

[3] H. H. Clark. *Using Language*. NewYork, NY: Cambridge University Press, 1996.

[4] D.E. Appelt and A. Kronfeld. "A computational model of referring". In *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, 640–647, 1987.

[5] S. E. Brennan and H. H. Clark. "Conceptual pacts and lexical choice in conversation". *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 1482–1493, 1996.

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

[7] M. J., Pickering and S. Garrod. "Toward a mechanistic psychology of dialogue". *Behavioral and brain sciences*, *27*(02), 169-190, 2004.

[8] D. A. Balota, J. E. Boland and L. W. Shield. "Priming in pronunciation: Beyond pattern recognition and onset latency". *Journal of Memory and Language*, 28(1), 14–36, 1989.

[9] J. M. Kahn and J.E. Arnold. "A processing-centered look at the contribution of givenness to durational reduction". *Journal of Memory and Language*, *67*(3), 311-325, 2012.

[10] D. G. Watson, J. E. Arnold and M.K. Tanenhaus. "Tic Tac TOE: Effects of predictability and importance on acoustic prominence in language production". *Cognition*, 106, 1548–1557, 2012.

[11] M. K. K. Halliday. *Intonation and grammar in British English*. The Hague: Mouton, 1967.

[12] C. Féry. "Intonation of focus in French". In C. Féry and W. Sternefeld (Eds.), *Audiatur Vox Sapientes: A Festschrift for Arnim von Stechow,* (Berlin: Akademi Verlag), 153-181, 2001.

[13] M. Dohen and H. Lœvenbruck. "Pre-focal rephrasing, focal enhancement and postfocal deaccentuation in French". In *Proceedings of Interspeech,* 2004.

[14] A. Michelas, C. Faget, C., Portes, A.-S. Lienhart, L. Boyer, C. Lançon, and M. Champagne-Lavau. "Do patients with schizophrenia use prosody to encode contrastive discourse status?". *Frontiers in Psychology*, 5:755, 2014.

[15] S-A. Jun and C. Fougeron. "Realizations of accentual phrase in French". *Probus 14*, 147-172, 2002.

[16] P. Welby. "French intonational structure: Evidence from tonal alignment". *Journal of Phonetics* 34(3), 343–371, 2006.

[17] A. Michelas and M. D'Imperio. "When syntax meets prosody: Tonal and duration variability in French Accentual Phrases". *Journal of Phonetics* 40(6), 816-829, 2012.

[18] E. Delais-Roussarie, B. Post, M. Avanzi, C. Buthke, A. Di Cristo, I. Feldhausen, S.-A. Jun, P. Martin, T. Meisenburg, A. Rialland, and R. Sichel-Bazin. "Intonational phonology of French: Developing a ToBI system for French". In S. Frota and P. Prietro (Eds.), *Intonational variation in Romance*. Oxford: OUP, 2016.

[19] D. R. Ladd. *Intonational phonology*. *Cambridge: Cambridge University Press,* 1996.

[20] D. Bates, M. Maechler and B. Dai. "lme4: linear mixed-effects models using S4 classes". R package version 0.9975-13, 2008.

[21] D.J. Barr, R. Levy, C. Scheepers, H.J Tily. "Random effects structure for confirmatory hypothesis testing: Keep it maximal". *Journal of Memory and Language 68(3)*, 255-278, 2013.

[22] R. H. Baayen, D.J., Davidson and D. M. Bates. "Mixed-effects modeling with crossed random effects for subjects and items". *Journal of memory and language*, *59*(4), 390-412, 2008.

[23] J. Vogels, E. Krahmer and A. Maes. "How cognitive load influences speakers' choice of referring expressions". *Cognitive science*, *39*(6), 1396-1418, 2015.

[24] J. German and M. D'Imperio. "The status of the initial rise as a marker of focus in French". *Language and Speech*, 59(2), 165-195, 2016.

[25] C. Beyssade, B. Hemforth, J.-M. Marandin and C. Portes. "Prosodic Markings of Information Focus in French". In H.-Y. Yoo and E. Delais-Roussarie (Eds.), *Actes d'Interfaces, Discours et Prosodie, 2009*, Paris, ISSN 2114-7612, 2009.

[26] E. G. Bard and M. P. Aylett. "Referential form, word duration, and modeling the listener in spoken dialogue". In J. C. Trueswell and M. K. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions* (pp. 173–191). Cambridge, MA: MIT Press, 2005.