Speech Prosody 2022
23-26 May 2022, Lisbon, Portugal

Prosody-to-Focus Mapping and Alternative Processing in Word Learning

Martin Ho Kwan Ip[1], *Alex de Carvalho*[2], *John Trueswell*[1]

[1]MindCORE, Integrated Language Sciences and Technology (ILST), University of Pennsylvania
[2]LaPsyDÉ, Centre National de la Recherche Scientifique (CNRS), Université de Paris
mhkip@sas.upenn.edu, alex.de-carvalho@u-paris.fr, trueswel@psych.upenn.edu

Abstract

The present experiment examined two issues concerning the role of prosody in word learning. First, we explored whether learners use prosodic focus to map words produced with contrastive stress onto contextually new visual referents. Second, we asked whether prosodic focus facilitates better memory for focused words and their contextual alternatives. In an eyetracking task, 48 monolingual English-speaking adults (18-40 yrs-old) were familiarized with videos of different people doing different actions. At test, participants saw two side-by-side videos, one showing a novel person performing a familiar action, and the other a familiar person performing a novel action. Participants then heard an utterance with prosodic focus on the noun or the verb (e.g., "*Now JOHNNY is blicking!*" vs. "*Now Johnny is BLICKING!*"). As predicted, participants paired the prosodically focused word with the contextually novel referent; participants who heard name-focused sentences looked longer at the novel person, while those who heard verb-focused sentences looked longer to the novel action. However, verb-focused sentences during word learning led to better recall of people's names than name-focused sentences. Further, eyegaze turns to alternative events during word learning led to better recall, indicating that eye movements during word learning support and reflect alternative processing.

Index Terms: prosodic focus, word learning, contextual alternatives, discourse-to-prosody mapping, memory

1. Introduction

How does prosody help listeners learn the meanings of words? In everyday conversations, listeners need to process not only the segmental content that determines what words they are hearing but also the suprasegmental structure that dictates *how* these words are produced. As Bolinger first pointed out decades ago, how a spoken word is perceived depends on its place above the lexical level, in the utterance intonation structure [1].

The goal of the present experiment is to examine how listeners use prosodic focus to infer and remember the meanings of accented novel words and their contextual alternatives. Despite growing research on the effects of prosodic boundary cues on word learning and segmentation across various languages and age groups [e.g., 2, 3, 4, 5, 6, 7], there is much less research on how listeners use prosodic focus cues to learn words in real time [c.f., 8, 9]. This is a major shortcoming because language learning is intricately tied to the discourse structure [10], and across many languages, prosody is the medium through which this discourse structure is expressed, albeit in many different ways (e.g., increased F_0 range, duration, and intensity in English and Mandarin [11, 12], only duration increase in Cantonese [13], accentual phrases in Korean [14]).

Importantly, prosodic focus may facilitate word learning by helping listeners map perceptually salient words onto referents that are contextually new or contrastive in the discourse. In English, these prosodically focused words, through specific pitch range movements and expansion, can affect the listener's attention patterns to relevant referents [15]. At the same time, listeners also perceive prosodically focused words for their contextual importance in the discourse structure; listeners not only process what is accented, as an acoustic cue, but also draw significance in meaning from what is not accented. For example, in a recent study, Spalek and Koch [16] presented listeners with utterances (e.g., "*Petra put fountain pens, notepads, and hole punchers in her bag*"), followed by another utterance with or without a prosodic focus (e.g., "*Petra used fountain pens/FOUNTAIN PENS")*. When later asked to recall the items (e.g., "*Which office supplies were there?"*), participants were better at remembering the contextual alternatives (i.e., notepads and hole punchers) if they previously heard the utterance with the prosodic focus. Thus, listeners not only perceive focus for its acoustic salience, but also for its importance in establishing a set of alternatives that are crucial for interpreting the intended meaning of an utterance [17, 18].

Although there have been many behavioral studies examining the role of prosodic focus in recall of contextual alternatives (e.g., [16, 19, 20, 21, 22]), only a handful of studies has explored this issue using a visual world paradigm, and to the best of our knowledge, there is no experiment looking at this effect using paradigms that involve live videos with novel words rather than static pictures or isolated sets of sentences. Moreover, it is also unclear if the effects of prosodic focus on word learning and recall of alternatives are influenced by the syntactic status of a word (e.g., nouns vs. verbs). For example, prior research in children suggests that verbs are harder to learn than nouns because processing verb meanings requires grasp of more advanced syntactic structure or conceptual mapping [23].

The experiment we report here forms part of a larger cross-language project examining prosodic focus and word learning in English- and French-speaking adults and toddlers. In the present English component, we aim to draw insights on word learning from adults by creating a context where they do not know the names of people or the words that describe what they are doing. Examining adults is crucial to understanding language acquisition, because it helps us probe the contextual factors that influence word learning and memory independent of the learner's conceptual abilities [see 24]. Participants will watch videos of people doing different actions while hearing utterances with focus on the nouns or verbs. We predict that participants will pay more attention to a new person if they hear a name-focused utterance, but to a new action if they hear a verb-focused utterance. Since focus affects recall of alternatives, we also predict that name-focused utterances will lead to better memory for names than verb-focused utterances.

10.21437/SpeechProsody.2022-3

2. Methods

2.1. Participants

Due to pandemic restrictions, all study sessions were conducted as a webcam-based experiment on Prolific. We tested 48 monolingual speakers of American English (aged between 18 to 40 years; 24 females). All participants were randomly assigned to either the noun accented or verb accented conditions (n = 24 in each condition). There was an equal number of men and women in each condition.

2.2. Design, Materials, and Procedures

We created a webcam-based eyetracking paradigm using PennController for Ibex [25]. The study was divided into two parts; (1) word learning and (2) word memory. In the word learning part, participants watched videos of people performing different actions (e.g., arm twirling) while listening to sentences produced by a female voice-over. In the word memory part, participants were asked to identify the names of the people they saw during word learning. The voice-over sentences were recorded by a native female speaker who produced them in child-friendly speech at a natural rate.

2.2.1. Part I: Word Learning

There were four word learning trials; two with boy pairs and two with girl pairs, all counterbalanced across participants and conditions. Each word learning trial contained three phases: warm-up, familiarization, and test (see Figure 1). Between each phase, we displayed a central fixation (i.e., a rainbow twirl) coupled with an attention-grabbing sentence (e.g., "*Oh look!*") to maintain participants' attention to the middle of the screen before the presentation of each video.

In the warm-up phase, participants saw a video of two people (i.e., two men or two women) standing next to each other waving at the camera. During this time, participants also heard the voice-over telling them the names of the two people in the screen (e.g., "*Oh look! It's Johnny and Billy! Do you see? It's Billy and Johnny.*"). However, the voice-over did not make clear which name referred to which person.

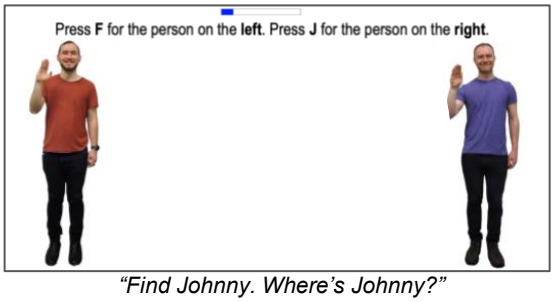
Immediately after the warm-up phase was the familiarization phase. Here, only one of the two people from the warm-up phase appeared on the screen, and s/he was performing an action, during which participants also heard a sentence that drew their attention to both the individual and the action (i.e., "*Wow! Look at him/her. What is s/he doing?*).

In the test phase, participants saw two side-by-side videos, one containing a familiar person performing an unfamiliar/novel action, and the other containing an unfamiliar/novel person performing a familiar action. Specifically, on one side of the screen, there was a video of the same person from the familiarization phase, this time performing a completely different novel action. On the other side, there was a video of a new person (i.e., the person who did not appear in the familiarization phase) performing the same action that participants saw during the familiarization phase. While watching these two side-by-side videos, participants heard the voice-over producing a test sentence that involved a noun (i.e., a person's name) and a verb (e.g., "*Now Johnny is blicking.*"). Participants in the noun accented condition heard test sentences where contrastive stress was placed on the noun (e.g., "*Now JOHNNY is blicking!*"), while participants in the verb accented condition heard prosodic stress on the verbs (e.g., "*Now Johnny is BLICKING!*"). The speaker who produced the voice-over sentences was asked to produce prosodic focus with a bitonal rising pitch accent, which has been shown to activate alternatives for accented words [20, 26].

Figure 1: *Word learning trial example.*

2.2.2. Part II: Word Memory

After going through the word learning trials, participants were presented with the memory trials (see Figure 2). Here, participants were presented with two side-by-side pictures of two of the people they saw from the word learning trials, and were asked by the voice-over to identify the people named (e.g., "*Where's Johnny? Find Johnny.*"). In response to the voice-over's commands, participants pressed either "F" (to select the picture on the left) or "J" (to select the picture on the right). We measured how well participants learned and remembered both the names that were being accented/referred to during the test phases of the word learning trials (e.g., "*Johnny*" from Figure 1), as well as their contextual alternatives (e.g., "*Billy*", the name of the other person). There were 16 memory trials.



"Find Johnny. Where's Johnny?"

Figure 2: *Word memory trial example.*

2.3. Stimuli Acoustic Analyses

Acoustic analyses of the test sentences were conducted based on simultaneous inspection of the waveform and the spectrogram in Praat [30]. The noun and verb tokens of the test sentences for each of the four word learning trials were annotated, and duration, mean F_0, maximum F_0, F_0 range, mean RMS-intensity, and maximum RMS-intensity were all measured (see Figure 5 for an example). The noun tokens across the accent conditions differed in duration, $t(3) = 4.42$, $p = .005$, mean F_0, $t(3) = 3.18$, $p = .019$, maximum F_0, $t(3) = 21.22$, $p < .001$, and F_0 range , $t(3) = 16.60$, $p < .001$, where in all cases the noun tokens showed greater duration and pitch in the noun accented condition than in the verb accented condition. The verb tokens in the verb accented condition had longer duration, $t(3) = 6.62$, $p < .001$, higher mean F0, $t(3) = 11.76$, $p < .001$, higher maximum F0, $t(3) = 52.32$, $p < .001$, and more expanded F0 range, $t(3) = 10.41$, $p < .001$, as well as greater mean intensity, $t(3) = 19.21$, $p < .001$, and maximum intensity, $t(3) = 17.44$, $p < .001$. These findings confirmed that prosodic focus occurred at their designated locations across the conditions.

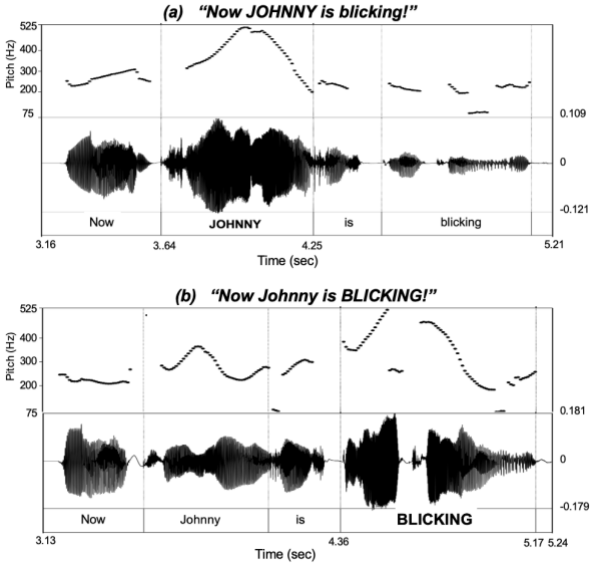


Figure 3: *Pitch contours and waveforms of an example. noun accented (a) versus verb accented (b) sentence pair.*

3. Results

3.1. Word Learning

All eyegaze data were analyzed manually by two annotators from the onset to the offset of the test phases (inter-rater reliability, $r = .98$, $p < .001$). Eyegaze to each side of the screen during the timecourse of the test phases were analyzed. Two areas of interest were defined, one comprising the left half of the screen, and the other comprising the right half of the screen.

We conducted a cluster-based permutation analysis [27], which has been used for previous eye-tracking studies (e.g., [28], [29]). Looking preference was calculated for 200 msec bins starting at the test phase onset. For each bin, a linear mixed effect regression model was conducted with focus condition (noun accented vs. verb accented) as fixed effects. The t-values for each coefficient were thresholded at 1.50. Clusters of above threshold values were identified, and the t-values were summed within each cluster. 1000 simulations randomly shuffling the accent conditions were conducted. For each simulation, the analysis calculated the size of the biggest cluster identified with the same procedure that was applied to the real data. A cluster of adjacent time points from the real data would show a significant effect of condition if the sum of the t-values in this particular cluster was greater than the highest t-value sum derived from clusters in 95% of the simulations, which ensures a p-value of 0.05. Results of the cluster-based permutations (see Figure 4) found a significant difference in looking time in a time window from 300 ms after the test phase onset until the end of the test phase (sum *t*-statistics = 752.91, $p < .001$).

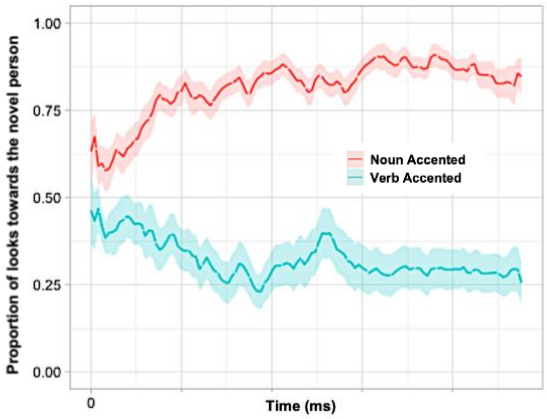


Figure 4: *Proportion of looks to the unfamiliar/novel person as a function of the noun vs. verb accented conditions, from the onset to the offset of the test phase. Note: 0.5 indicates no preference; +0.5 indicates preference for the novel person; -0.5 indicates preference for the novel action.*

Furthermore, averaging looking times across all the word learning trials (Figure 5) revealed that participants in the noun accented condition looked longer at the unfamiliar person than those in the verb accented condition, $t(46) = 7.88$, $p < .001$).

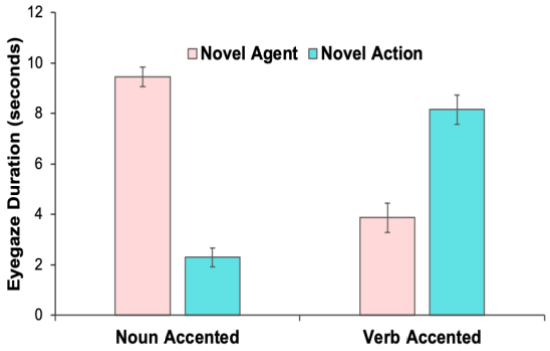


Figure 5: *Average overall looking times (in secs) to each of the videos presented in the test phase. Looks to the novel agent/person (pink) and to the novel action (blue) as a function of accent condition.*

3.2. Word Memory

Two-tailed independent-groups t-tests were conducted to compare participants' memory of people's names as a function of their accent condition. We also analyzed the effect of accent condition separately for men and women, because previous studies across different languages have found that the effect of prosodic focus improves recall only in women [e.g., 16, 19].

Overall, averaging across all the memory trials, participants who went through the verb accented condition during word learning had better recall accuracy for people's names (M = 75.78%, SD = 28.47%) than participants who went through the noun accented condition (M = 63.54%, SD = 17.43%), $t(46)$ = 2.67, p = .028 (see Figure 6). There was also a gender difference; only female participants showed better recall from being in the verb accented condition, $t(22)$ = 2.31, p = .031. For male participants, the result trend was in the same direction, albeit nonsignificant.

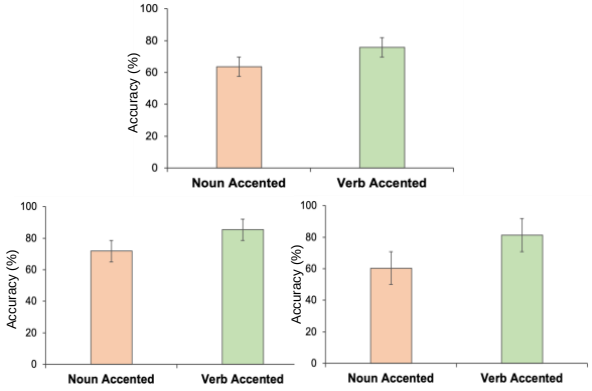


Figure 6: *Overall noun memory accuracy (top) and memory accuracy for accented names (bottom left) and contextual alternatives (bottom right).*

As already mentioned, we examined how well participants remembered both the names that were being referred to in the word learning trial test phases (e.g., "*Johnny*" in "*Now Johnny is blicking*", in the situation where both Johnny and Billy were present), as well as their contextual alternatives (e.g., their ability to remember "*Billy*", the unmentioned name during the test phase). Here, results only revealed significant effects of accent condition in memory trials where participants were shown pictures of two people who were not paired together during the word learning trials (e.g., asked to "*Find Johnny*" when being shown "Johnny" and "Kenny", rather than "Johnny and Billy", see Figure 2 for such example).

For the names that were being referred to in the word learning trial test sentences, we found that participants in the verb accented condition (M = 71.88%, SD = 25.87%) had marginally better recall than participants in the noun accented condition (M = 85.42%, SD = 23.22%), $t(46)$ = 1.91, p = .063. However, there was, again, gender differences, in which only the female participants showed a significant effect of the accent condition of the word learning trials, $t(22)$ = 2.93, p = .008.

For the names that were contextual alternatives, analyses revealed a significant effect of accent condition; again, participants who went through the verb accented condition during word learning had higher recall accuracy (M = 81.25%, SD = 21.18%) than participants who went through the noun accented condition (M = 60.40%, SD = 28.47%), $t(46)$ = 2.88, p = .006. However, this time, only the male participants showed the significant effect of accent condition, $t(22)$ = 2.67, p = .014.

3.3. Correlational Analyses

Additional analyses were conducted to examine whether there was any association between participants' eyegaze patterns during the word learning trials and their word memory. For the participants who went through the verb accented condition during word learning, analyses revealed a significant positive correlation between participants' eyegaze duration at novel people doing familiar actions and their memory for people's names, r = .05, p = .004.

4. Discussion

The present experiment examined how prosodic focus supports word learning and memory. By adopting a visual world paradigm involving live videos of people and actions with varying levels of familiarity, we show that listeners could map prosodically focused words onto contextually new and contrasted referents during word learning. However, contrary to our hypotheses, participants who heard the verb focus had better recall accuracy for people's names, than the participants who heard the noun/name focus. These memory findings are intriguing and somewhat puzzling, because while they show an effect of focus on recall improvement, it was in the opposite direction from our predictions.

One reason for such a finding could be the different syntactic status of the nouns vs. verbs. As mentioned earlier, verbs may be harder to process and learn than nouns [23]. Given that processing difficulty generally leads to better learning [31], participants in the verb accent condition might have, overall, learned and remembered all the words better because the sentences with verb focus were harder to process/interpret. At the same time, further analyses indicated that alternative processing was still at play in listeners' processing of contextual alternatives. In our correlational analyses, there was a positive association between memory accuracy for people's names and the degree to which participants hearing verb focus during word learning switched their gaze to look at the novel person. Note that when hearing verb focus, participants did look longer at the novel action performed by the familiar person, compared to the familiar action performed by a novel person. However, many participants have switched their gaze a few times to glance at the novel person, and the degree to which they looked at this alternative event was correlated with their recall memory of people's names. These findings indicate that eyegaze towards alternative events might have been responsible for facilitating listeners' processing of contextual alternatives in real time.

Work in our lab is currently underway to examine the developmental origins of infants' use of prosodic focus in word learning and alternative processing. A key question is whether all human infants have a developmentally early ability to map prosodically focused words (e.g., increased pitch range) onto their discourse referents, even in a language where speakers are more likely to use non-prosodic means to mark focus (e.g., clefting in French [32]). From a methodological standpoint, the present eye-tracking experiment introduces a useful new way to test listeners' prosodic sensitivity in language learning.

5. Acknowledgements

We acknowledge support from Penn's ILST Initiative. We are grateful to members of the Trueswell-Gleitman and Swingley labs for their comments. We also thank Abimael Hernandez for technical assistance, and Steven-John Kounoupis and Sebleh Alfa for their help in the eyegaze annotations.

6. References

[1] D. L. Bolinger, “Around the edge of language: Intonation,” *Harvard Educational Review*, vol. 34, no. 2, pp. 282–296, 1964.

[2] A. de Carvalho, A. X. He, J. Lidz, and A. Christophe, “Prosody and function words cue the acquisition of word meanings in 18-month-old Infants,” *Psychological Science*, vol. 30, no. 3, pp. 319–332, Mar. 2019, doi: 10.1177/0956797618814131.

[3] A. Christophe, S. Peperkamp, C. Pallier, E. Block and J. Mehler, "Phonological phrase boundaries constrain lexical access I. adult data," *Journal of Memory and Language*, vol. 51, no. 4, pp. 523-547, Nov. 1, 2004.

[4] A. Christophe, J. Mehler and N. Sebastián-Gallés, "Perception of prosodic boundary correlates by newborn infants," *Infancy*, vol. 2, no. 3, pp. 385-394, Jul. 2001.

[5] S. Kim, M. Broersma and T. Cho, "The use of prosodic cues in learning new words in an unfamiliar language," *Studies in Second Language Acquisition*, vol. 34, no. 3, pp. 415-444, Aug. 2012.

[6] A. Christophe, E. Dupoux, J. Bertoncini and J. Mehler, "Do infants perceive word boundaries? an empirical study of the bootstrapping of lexical acquisition," *The Journal of the Acoustical Society of America*, vol. 95, no. 3, pp. 1570-1580, Mar. 1994.

[7] M. Shukla, K.S. White and R.N. Aslin, "Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants," *Proceedings of the National Academy of Sciences*, vol. 108, no. 15, pp. 6038-6043, Apr. 2011.

[8] J.P.M. Fikkert and A.J. Chen, "The role of word-stress and intonation in word recognition in Dutch 14- and 24-month-olds," in *Proceedings of BUCCLD*, Boston, USA, Nov. 2011, pp. 222-232.

[9] S. Grassmann and M. Tomasello, "Prosodic stress on a word directs 24-month-olds’ attention to a contextually new referent," *Journal of Pragmatics*, vol. 42, no. 11, pp. 3098-3105 , Nov. 2010.

[10] C. Fisher, K. Jin and R.M. Scott, "The developmental origins of syntactic bootstrapping," *Topics in Cognitive Science*, vol. 12, no. 1, pp. 48-77, Aug. 2020.

[11] Y. Chen and C. Gussenhoven, "Emphasis and tonal implementation in standard chinese," *Journal of Phonetics*, vol. 36, no. 4, pp. 724-746, Oct. 2008.

[12] Y. Xu, "Effects of tone and focus on the formation and alignment of f0contours," *Journal of Phonetics*, vol. 27, no. 1, pp. 55-105, Jan. 1999.

[13] H.S.H. Fung and P.P.K. Mok, "Temporal coordination between focus prosody and pointing gestures in Cantonese," *Journal of Phonetics*, vol. 71, pp. 113-125, Nov. 2018.

[14] S. Jun, "The accentual phrase in the korean prosodic hierarchy," *Phonology*, vol. 15, no. 2, pp. 189-226, Dec. 1998.

[15] J. C. Thorson and J. L. Morgan, “The role of intonation in early word recognition and learning,” in *Proceedings of Speech Prosody*, Dublin, Ireland, pp. 1159-1163, May. 2014.

[16] X. Koch and K. Spalek, "Contrastive intonation effects on word recall for information-structural alternatives across the sexes," *Memory and Cognition*, vol. 49, no. 7, pp. 1312-1333, Oct. 2021.

[17] M. Rooth, “A theory f focus interpretation,” *Natural Language Semantics*, vol. 1, pp. 75-116, Feb. 1992.

[18] M. Krifka and R. Musan, “Information structure: Overview and linguistic issues,” in *The Expression of Information Structure*, M. Krifka and R. Musan (eds.). Berlin: Mouton de Gruyter, 2012, pp. 1-44.

[19] A. Tjuka, H.T.T. Nguyen and K. Spalek, "Foxes, deer, and hedgehogs: The recall of focus alternatives in vietnamese," *Laboratory Phonology*, vol. 11, no. 1, pp. 16, Oct. 2020.

[20] S.H. Fraundorf, D.G. Watson and A.S. Benjamin, "Recognition memory reveals just how CONTRASTIVE contrastive accenting really is," *Journal of Memory and Language*, vol. 63, no. 3, pp. 367-386, Oct. 2010.

[21] E.M. Husband and F. Ferreira, "The role of selection in the comprehension of focus alternatives," *Language, Cognition and Neuroscience*, vol. 31, no. 2, pp. 217-235, Feb. 2016.

[22] B. Braun and L. Tagliapietra, "The role of contrastive intonation contours in the retrieval of contextual alternatives," *Language and Cognitive Processes*, vol. 25, no. 7-9, pp. 1024-1043, Sept. 2010.

[23] L. Gleitman, “The structure of verb meanings,” *Language Acquisition*, vol. 1, no. 1, pp. 3-55, Jan. 1990.

[24] L.R. Gleitman, K. Cassidy, R. Nappa, A. Papafragou and J.C. Trueswell, "Hard words," *Language Learning and Development*, vol. 1, no. 1, pp. 23-64, Jan. 2005.

[25] J. Zehr and F. Schwartz, “PennController for internet based experiments (IBEX). 2008.

[26] J. B. Pierrehumbert and J. Hirschberg, “The meaning of intonational contours in the interpretation of discourse,” in *Intentions in Communication*, P. Cohen, J. Morgan, and M. Pollack (eds.). Cambridge, MA: MIT Press, 1990, 271-311.

[27] E. Maris and R. Oostenveld, "Nonparametric statistical testing of EEG- and MEG-data," *Journal of Neuroscience Methods*, vol. 164, no. 1, pp. 177-190, Aug. 2007.

[28] A. de Carvalho, M. Babineau, J.C. Trueswell, S.R. Waxman and A. Christophe, "Studying the real-time interpretation of novel noun and verb meanings in young children," *Frontiers in Psychology*, vol. 10, Feb. 2019.

[29] B. Ferguson, E. Graf and S.R. Waxman, "When veps cry: Two-year-olds efficiently learn novel words from linguistic contexts alone," *Language Learning and Development*, vol. 14, no. 1, pp. 1-12, May. 2018.

[30] P. Boersma and D. J. M. Weenink, “Praat, a system for doing phonetics by computer,” *Glot International*, vol. 5, no. 9/10, pp. 1381–3439, Jan. 2001.

[31] V.I. Schneider, A.F. Healy and L.E. Bourne Jr., "What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer," *Journal of Memory and Language*, vol. 46, no. 2, pp. 419-440, Feb. 2002.

[32] C. Féry, “Focus and phrasing in French,” in *Audiatur Vox Sapientiae*, C. Féry and W. Sternfeld (eds.). Berlin: Akademie Verlag, 2001, pp. 153-181.