Probabilistic prosody: Effects of relative speech rate on perception of (a) word(s) several syllables earlier

Meredith Brown[1], *Laura C. Dilley*[2], *Michael K. Tanenhaus*[1]

[1]Department of Brain & Cognitive Sciences, University of Rochester, Rochester, NY, US
[2]Department of Communicative Sciences & Disorders, Michigan State University, East Lansing, MI, US

mbrown@bcs.rochester.edu, ldilley@msu.edu, mtan@bcs.rochester.edu

Abstract

Speech perception depends on the ability to rapidly accommodate considerable variability in speech rate. We present results from two eye-tracking experiments indicating that listeners use context speech rate to generate, maintain, and update probabilistic hypotheses about the timing and number of constituents in upcoming speech. Participants heard utterances containing polysyllabic nouns preceded by indefinite articles and followed by [s]-initial words (e.g. *...saw a raccoon slowly...*). We altered the speech rate of the indefinite article and of the [s] with respect to surrounding context, manipulating the likelihood that the item would be perceived as singular (*a raccoon*) vs. plural (*raccoons*). Shorter indefinite articles elicited higher proportions of fixations to plural target pictures than longer articles both before and after the processing of [s], demonstrating that listeners made rapid use of prosodic cues to the presence or absence of the article. Importantly, fixations were also influenced by the duration of [s] relative to context speech rate. These findings suggest that listeners maintain and update provisional speech-rate hypotheses across multiple morphophonemic units. We interpret these results with respect to probabilistic approaches to spoken language understanding.

Index Terms: perception of prosody, speech rate, expectations, eye movements, language comprehension

1. Introduction

The realization of prosodic information in speech, such as pitch accents, speech tempo, and other intonational features, is highly variable (e.g. [1], [2], [3]). This variability poses numerous challenges for spoken language processing. For example, the realization of temporal speech cues like voice onset time depends on an individual's overall speech rate. Comprehension therefore crucially depends on the ability of listeners to interpret prosodic cues and rate-dependent speech cues with respect to surrounding context (e.g. [4], [5], [6]). A comprehensive understanding of the role of prosody in spoken language comprehension requires an explanation of how listeners accommodate contextual information during real-time processing.

In this paper we explore a possible explanation for how listeners interpret prosody in context based on emerging *data-explanation* approaches to perception and cognition. Data-explanation approaches posit a central role for *generative processes* within perceptual systems that give rise to probabilistic expectations about incoming sensory input based on high-level representations and contextual information (e.g. [7], [8]). During speech perception, we hypothesize that listeners continuously make and update inferences about the source of the speech signal (i.e. the communicative intention of the speaker) by comparing internally generated probabilistic expectations about the acoustic realization of the speech signal to the actual speech signal as it unfolds. This provides a potential explanatory framework for the integration of multiple distinct and temporally distributed constraints during spoken language processing [9]. Particularly compelling from this perspective are so-called *distal prosody* effects [10], [11]. For example, manipulating pitch and timing patterns across utterance material several syllables before a temporarily ambiguous word (e.g. *panda*, which can initially be interpreted as *pan*) influences the time course of lexical competition, even when the prosodic characteristics of the target word itself are unaltered [12], [13]. These findings suggest that listeners develop expectations based on prosodic patterns in speech that extend across a relatively wide window.

The data-explanation framework provides a potential explanation for how expectations based on preceding prosody are mapped onto the acoustic-phonetic properties of the unfolding speech signal. It also makes the prediction that listeners continuously update their provisional hypotheses about the source of the speech signal on the basis of additional downstream information. The present study investigates this prediction by capitalizing on effects of context speech rate on the number of words that are perceived within a stretch of speech [11]. Compressing the speech rate of portions of an utterance surrounding a highly coarticulated word (e.g. the underlined segments surrounding the determiner "a" in *The Smiths wouldn't buy a Butterball...*) reduces the likelihood that listeners report hearing this word. Likewise, when the word in question is not present in the signal (e.g. *buy Butterball...*), slowing down the segments surrounding its potential location makes listeners more likely to perceive the word within the slow portion of speech.

In the present study, we investigate whether and how listeners maintain and update provisional prosodic percepts based on downstream cues, by manipulating the relative speech rate of multiple temporally-distributed cues to the plurality of a noun phrase – the presence or absence of the indefinite determiner *a* before the noun and of the plural marker *–s* following the noun. Our experiments use the *visual world paradigm* to examine listeners' eye movements to pictures depicting singular and plural versions of the target noun phrase, providing an index of the time course of interpretation [14], [15]. The goals of this work were to determine (a) whether and how listeners combine multiple temporally distributed prosodic cues to the plurality of a referring expression; and (b) whether the time course of prosodic cue integration is consistent with the hypothesis that listeners rapidly update their previous prosodic expectations and provisional percepts in light of downstream information.

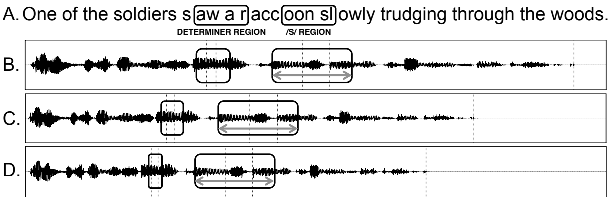

Figure 1: *Illustration of the distal /s/ speech rate manipulation in Experiment 1: (a) Example stimulus sentence; (b) 86% determiner, 95% distal /s/; (c) 86% determiner, 75% distal /s/; (d) 86% determiner, 65% distal /s/. Boxes indicate the position and duration of determiner and /s/ regions, respectively.*

2. Experiment 1

2.1. Methods

2.1.1. Participants

We recruited 36 students from the University of Rochester to participate in Experiment 1. All participants were native speakers of American English and had normal hearing and normal or corrected-to-normal visual acuity.

2.1.2. Materials

The stimuli were 54 grammatical declarative sentences containing a highly imageable singular noun phrase (e.g. *a raccoon*) followed by a word starting with an /s/ (e.g. Fig. 1a). Recordings of these sentences were elicited from 12 speakers. These speakers also recorded 81 filler items whose target nouns were equally likely to be singular or plural and whose number was unambiguous (e.g. *those bathrobes*). We selected one token of each item for use in the experiment, consisting of 4-5 critical item tokens and 6-7 filler item tokens from each speaker. Critical item tokens were selected such that they had a relatively high degree of coarticulation on the indefinite determiner "a" and continuous articulation of the target word and the following /s/, such that the presence or absence of the plural marker *–s* was not clearly signaled.

Stimuli were manipulated in two ways (Fig. 1b-d). First, the speech rate of the region consisting of the determiner and the segments immediately surrounding it (*determiner region*) was compressed to 92%, 86%, or 78% of its original rate. Then, we compressed the speech rate of utterance context preceding and following the segments surrounding the /s/ following the target word (*/s/ region*), such that it was 95%, 75%, or 65% of its original rate. The absolute physical duration of the /s/ region remained the same across conditions, but its speech rate relative to surrounding context increased with successive levels of context speech rate compression. Levels of each manipulation were selected based on norming data. The global speech rate of filler items was manipulated such that equal numbers of items were compressed to 95%, 75%, or 65% of their original rate.

2.1.3. Procedure

On each trial, participants were presented with a computer screen containing a four-picture visual display. The display contained singular and plural versions of the target word and of a distractor picture. Each participant heard a single version of each item over Sennheiser HD 570 headphones after 500 ms of display preview. Their task was to click on the picture that they heard referred to in each sentence. Eye movements were recorded using a head-mounted SR Research EyeLink II system sampling at 250 Hz, with drift correction procedures performed following every fifth trial.

Two lists were created by pseudo-randomizing trial order and rotating picture positions 180 degrees. An additional set of two lists was created by reversing the order of these lists. An equal number of critical items in each list were assigned to each of the nine pairings of determiner condition and distal speech rate condition. The assignment of items to conditions was counterbalanced across participants. All lists started with six filler items to ensure that participants were familiar with the task prior to encountering critical items.

2.1.4. Analyses

Response choices and fixations were analyzed separately. Data from trials on which the participant incorrectly selected one of the two distractor pictures (less than 0.5% of trials) were excluded. Selections of singular vs. plural target pictures were analyzed using multilevel logistic regression. Proportions of fixations to singular and plural target pictures on each trial were averaged across two windows of interest: (a) an *early window* 400 ms in duration (the mean duration of the target word), starting 200 ms before and ending 200 ms after the onset of the /s/; and (b) a *late window* between 200–1000 ms following the onset of the /s/. Both windows were selected to take into account a 200 ms delay for programming and executing fixations. Mean proportions of fixations to the plural target picture were divided by the mean propotion of fixations to both target pictures to calculate a *plural target advantage ratio* across each window for each trial, which was transformed using the empirical logit function [16], [17]. Plural target advantage ratios were analyzed using linear regression. The significance of predictors in the linear regression models was estimated by assuming convergence of the t distribution with the z distribution [18]. Models were computed using the *lme4* package in R (version 2.15.0) [19], [20]. All regression models had determiner speech rate, distal speech rate, and their interactions as fixed effects, and full random effects structure except as noted due to lack of convergence [21]. Factors were contrast coded with the least rate-manipulated level set as the reference level (i.e. 92% determiner, 95% distal /s/). Model comparison procedures were used to remove fixed effects that did not contribute significantly to model fit according to the likelihood ratio test [18].

2.2. Predictions

We predicted that compressing determiner speech rate would result in increased selections of plural target pictures and higher plural target advantage ratios in both the early and late windows, replicating early effects of relative speech rate on spoken language processing observed in previous work [22]. Importantly, we predicted that compressing speech rate of material distal to the /s/ (effectively slowing down the speech rate of /s/ relative to surrounding context) would also result in increased plural target picture choices and proportions of fixations to plural target pictures following the onset of /s/.

2.3. Results and discussion

2.3.1. Picture choices

Participants' response choices were consistent with our predictions (Fig. 2). The multilevel logistic regression model of response choices confirmed that participants were more likely to select plural target pictures when the determiner had a faster

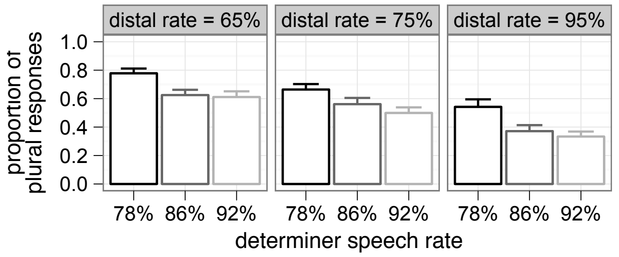

Figure 2: *Proportions of plural responses in Experiment 1.*

speech rate (66% plural responses for the most compressed determiner compared to 48% for the least compressed determiner; β=1.42, z=5.76, p<.0001). The distal /s/ manipulation also had the predicted effect. Participants were most likely to choose plural target pictures when the /s/ was surrounded by the most compressed speech, and therefore had a slower speech rate with respect to the utterance context (67% plural responses). They were least likely to choose plural target pictures when the /s/ was surrounded by the least compressed speech (42% plural responses; β=1.89, z=7.96, p<.0001).

2.3.2. *Proportions of fixations*

Figure 3 shows fixations to plural and singular target pictures over time with respect to the onset of the /s/. Analysis of plural target advantage ratios across this window revealed early effects of determiner speech rate on fixation proportions (Fig. 3, top). More compressed determiners were associated with more looks to plural target pictures and fewer looks to singular target pictures, compared to less compressed determiners (β=0.11, t=1.92, p_{est}=.055). This finding replicates early effects of speech rate on determiner perception previously observed in related work and suggests that effects of speech rate manipulation on the perception of short words have a locus in perceptual expectations [22]. Effects of determiner speech rate persisted into the late analysis window (β=.13, t=3.97, p_{est}<.0001).

The distal /s/ speech rate manipulation also exhibited the predicted effects within the late analysis window, during and following the processing of the /s/ (Fig. 3, bottom)[1]. Plural target advantage ratios were highest when the /s/ was surrounded by the most compressed speech, and therefore had a slower speech rate than the utterance context, than when the /s/ was surrounded by relatively slow speech (β=.17, t=3.61, p_{est}<.0005).

These results suggest that determiner perception is influenced by prosodic information influencing whether the /s/ multiple syllables downstream is perceived as containing the plural morpheme *–s*. However, it is also possible that these effects are instead attributable to the absolute duration of the determiner. The determiner is compressed first as part of the determiner region manipulation, and then as part of the distal /s/ speech rate manipulation (which involves compressing all material distal to /s/, including the determiner). The absolute duration of the determiner therefore covaries with the relative speech rate of the /s/. For example, in the 86% determiner speech rate condition, the absolute duration of the determiner is in fact 82% of its initial duration in the slowest distal speech rate condition and 56% in the fastest (cf. Fig. 1b-d). It is therefore possible that apparent effects of the speech rate of /s/ relative to surrounding context could instead have resulted from effects of distal speech rate compression on the absolute duration of the determiner. These possibilities cannot be distinguished on the basis of the time course data, because of the overall effects of speech rate compression on fixations prior to the onset of the target noun phrase.

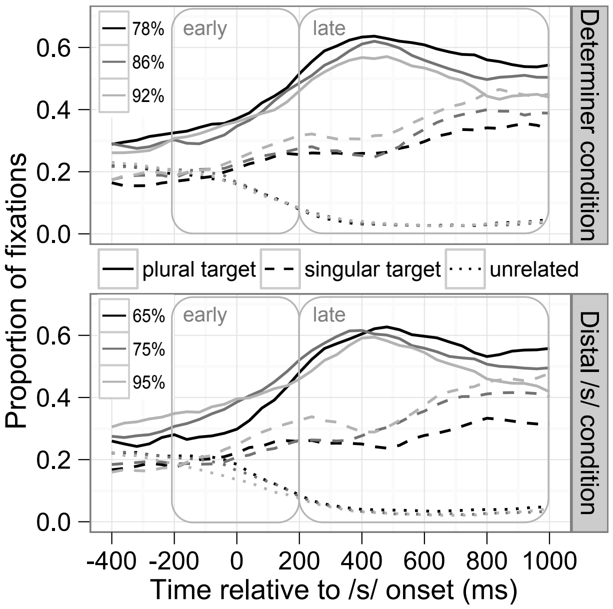

Figure 3: *Proportions of fixations in Experiment 1, by determiner (top) and distal /s/ (bottom) conditions. Superimposed shapes depict early and late analysis windows.*

To address this potential confound, we conducted a second visual world experiment in which we manipulated the speech rate of the /s/ by slowing down the /s/ region itself, rather than by compressing surrounding context. This proximal /s/ speech rate manipulation had no effect on the absolute duration of the determiner. Thus, if the effects that we observed in Experiment 1 were merely attributable to effects of the distal speech rate manipulation on the absolute duration of the determiner, we would not expect to observe effects of the proximal /s/ speech rate manipulation on picture choices or fixations to singular or plural pictures. If, however, the effects found in Experiment 1 were due to the speech rate of the /s/ relative to surrounding context, we would expect to see effects of the proximal /s/ manipulation that are similar to those that we observed in Experiment 1.

3. Experiment 2

3.1. Methods

Participants were 36 University of Rochester students meeting the same criteria as for Experiment 1. In addition, the experi-

[1] We also found marginally significant effects of the distal /s/ speech rate manipulation on fixations within the early window (i.e. before the processing of the /s/; β=.10, t=1.74, p_{est}=.082). Post-hoc analyses revealed that the logit-transformed sum of fixations to both target pictures differed as a function of distal speech rate condition (β=-0.17, t=-3.89, p_{est}<.0005). Increased compression of the utterance context was associated with lower proportions of fixations to target pictures within the early analysis window, because the analysis window was time-locked to the onset of the /s/. It is therefore likely that these effects of distal /s/ speech rate manipulation on pre-/s/ fixations are merely attributable to information about the target word becoming available at different times and different rates. Importantly, because the baseline effects observed in the early analysis window were in the opposite direction of the predicted effects of distal speech rate following the onset of the /s/, they did not complicate interpretation of effects in the later analysis window.

ment setup, stimulus lists, data collection, and analysis procedures were the same as in Experiment 1.

We used the recordings from Experiment 1 to create the stimuli for Experiment 2. The determiner manipulation was the same as in Experiment 1. However, instead of manipulating the relative speech rate of /s/ by compressing the speech rate of material surrounding the /s/ region, we instead manipulated the /s/ region directly by slowing its speech rate to 110%, 150%, or 170% of its original rate. Following this proximal /s/ speech rate manipulation, all critical and filler items were globally compressed to 85% of their original duration, to maintain similarity with Experiment 1 in terms of global stimulus characteristics and overall experiment duration.

3.2. Results and discussion

3.2.1. *Picture choices*

The multilevel logistic regression model of response choices. Participants again selected plural target pictures more frequently when the determiner region had a faster speech rate (65% in the fastest condition compared to 46% in the slowest condition; β=1.43, z=6.86, p<.0001). Crucially, the speech rate of /s/ also influenced response choices in the predicted direction, such that participants were more likely to select plural pictures when the /s/ region had a slower speech rate (62% plural responses in the slowest condition compared to 43% in the fastest condition; β=1.48, z=8.76, p<.0001). This suggests that the effects of the speech rate of /s/ relative to surrounding context influenced listeners' judgments in both experiments, rather than simply the absolute duration of the determiner.

3.2.2. *Proportions of fixations*

The multilevel linear regression of plural target advantage ratios in the early analysis window indicated no significant effects of determiner or /s/ speech rate manipulation. The determiner manipulation used in these experiments was subtle relative to manipulations used in previous related work, in which early effects of determiner speech rate were reliably observed (Brown et al., 2012). Previous work also used a multi-word target expression, providing a larger window of analysis with more statistical power. Although the manipulation we used in the present experiments elicited robust effects in response choices and in overall fixation behavior, it is possible that it was nevertheless too subtle to elicit large enough effects to be reliably observed immediately after the processing of the determiner, even though numerical trends in the predicted direction emerged within this window. In addition, other factors such as a lack of variation in the global speech rate of filler items may have contributed to subtle differences in time course and/or magnitude of effects across experiments.

Crucially, however, analysis of plural target advantage ratios during the late analysis window revealed significant effects of not only determiner but also /s/ speech rate. As predicted, more compressed determiners were associated with higher plural target advantage ratios (i.e. more fixations to plural pictures and fewer to singular pictures; β=.28, t=5.80, p_{est}<.0001). In addition, plural target advantage ratios were higher when the speech rate across the /s/ region was the slowest (β=.21, t=4.33, p_{est}<.0001). This finding, together with the significant effect of /s/ speech rate on picture choices, indicates that the effects of Experiment 1 cannot be explained merely on the basis of the absolute duration of the determiner across different levels of the distal /s/ speech rate manipulation. Eliminating this confound provides stronger evidence that the interpretation of function words can be modulated by information encountered considerably later in the utterance.

4. Discussion

Our results provide evidence that listeners combine multiple temporally distributed prosodic cues to the plurality of a referring expression during real-time spoken language comprehension. Prosodically conditioned percepts of short words like determiners can be influenced by congruent or conflicting information in the speech signal that occurs substantially downstream. These findings suggest that listeners maintain and update provisional hypotheses about previously encountered material across multiple morphophonemic units.

Our findings are closely aligned with recent work demonstrating that language processing is influenced by information spanning a wider temporal integration window than standardly assumed [23], [24], [25]. For example, when listeners hear a target word *leash* in a sentence context, lexical competition with *leaves* is stronger when the target word follows the verb *shakes* (whose rhyme *rakes* is semantically related to *leaves*) than when it follows the verb *rattles* [23]. This suggests that residual uncertainty about the perceived verb influences lexical competition effects several syllables downstream.

The present work further demonstrates that these "right-context" effects also extend to uncertainty about the timing and number of constituents in preceding speech, based on preceding prosodic information. These findings are difficult to explain with respect to traditional feed-forward models of spoken language comprehension that assume that listeners map acoustic patterns onto more abstract representations (e.g. words and morphemes) prior to interpreting the perceived representations with respect to sentence- and discourse-level context. Rather, our results suggest that listeners maintain and update probabilistic inferences about speakers' intended meaning (such as the intention to produce a singular or plural construction) based on available prosodic information across a relatively wide window.

5. Conclusions

Prosody influences spoken language processing in a gradient, probabilistic fashion. Further, prosodically-conditioned percepts are maintained and updated across multiple morphophonemic units. These findings are most naturally explained within a probabilistic data-explanation account of spoken language processing, involving probabilistic inference about the communicative intentions that give rise to the acoustic realization of an utterance. These inferences inform fine-grained probabilistic expectations about how aspects of lexical alternatives will be realized in context. They can be also updated in light of subsequent information encountered in the unfolding utterance.

6. Acknowledgements

This work was supported by an NSF predoctoral fellowship to MB, NSF grant BCS-0847653 to LCD, and NIH grants HD073890 and HD027206 to MKT. We gratefully thank Dana Subik and Chelsea Marsh for assistance with recruiting and testing participants, and the audiences at AMLaP 2013 and AP-CAM 2013 for feedback on earlier versions of this work.

7. References

[1] Ladd, D. R. (2008). *Intonational phonology*, (2nd Ed), Cambridge Studies in Linguistics.

[2] Badino, L. & Clark, R. A. J. (2007). Issues of optionality in pitch accent placement. In *Proc. 6th ISCA Speech Synthesis Workshop*, Bonn, Germany, 2007.

[3] Jacewicz, E., Fox, R. A., & Wei, L. (2010). Between-speaker and within-speaker variation in speech tempo of American English. *The Journal of the Acoustical Society of America, 128*, 839-850.

[4] Miller, J. (1987). Rate-dependent processing in speech perception. In A. Ellis (ed.), *Progress in the psychology of language* (pp. 119-157). London: Erlbaum Associates.

[5] Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. (1978). Perceptual integration of acoustic cues for stop, fricative, and affricate manner. *Journal of Experimental Psychology: Human Perception and Performance, 4(4)*, 621-637.

[6] Reinisch, E., Jesse, A., & McQueen, J. M. (2011). Speaking rate from proximal and distal contexts is used during word segmentation. *Journal of Experimental Psychology: Human Perception and Performance, 37(3)*, 978-996.

[7] Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences, 11(10)*, 428-434.

[8] Clark, A. (2013). Whatever next? Predictive brains, situated agents and the future of cognitive science. *Brain and Behavioral Sciences, 36,* 181-204.

[9] Farmer, T. A., Brown, M., & Tanenhaus, M. K. (2013). Prediction, explanation, and the role of generative models in language processing [Commentary]. *Behavioral and Brain Sciences, 36,*, 31-32.

[10] Dilley, L., & McAuley, J. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language, 59,* 291-311.

[11] Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science, 21(11),* 1664-1670.

[12] Brown, M., Salverda, A. P., Dilley, L. C., & Tanenhaus, M. K. (2011). Expectations from preceding prosody influence segmentation in online sentence processing. *Psychonomic Bulletin and Review, 18(6),* 1189-1196.

[13] Brown, M., Salverda, A. P., Dilley, L. C., & Tanenhaus, M. K. (under review). Metrical expectations from preceding prosody influence spoken-word recognition. *Journal of Experimental Psychology: Human Perception and Performance.*

[14] Cooper, R. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology, 6*, 84-107.

[15] Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268,* 1632-1634.

[16] Barr, D. J. (2008). Analyzing "visual world" eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*, 457-474.

[17] Cox, D. R. (1970). *The analysis of binary data.* London: Chapman and Hall.

[18] Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed- effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390-412.

[19] Bates, D., Maechler, M., & Bolker, B. (2011). lme4: Linear mixed-effects models using S4 classes (R package, version 0.999375-42) [Computer software]. Online: http://CRAN.R-project.org/package=lme4.

[20] R Development Core Team. (2012). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Online: http://www.R-project.org.

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

[22] Brown, M., Dilley, L. C., & Tanenhaus, M. K. (2012). Real-time expectations based on context speech rate can cause words to appear or disappear. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1374-79).

[23] Johnstone, S., Trueswell, J., & Dahan, D. (2013). Partially activated words participate in combinatory semantic interpretation during sentence processing. Talk presented at the 26th CUNY Conference on Human Sentence Processing (Columbia, SC).

[24] Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences, 106*, 21086-2109.

[25] McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). Within-category VOT affects recovery from "lexical" garden-paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language, 60(1),* 65-91.