Flexible Use of Phonological and Visual Memory in Language-mediated Visual Search

Daniel F. Pontillo (dpontillo@bcs.rochester.edu)
Anne Pier Salverda (asalverda@bcs.rochester.edu)
Michael K. Tanenhaus (mtan@bcs.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester
Meliora Hall, Box 270268, Rochester, NY 14627-0268

Abstract

In language-mediated visual search, memory and attentional resources must be allocated to simultaneously process verbal instructions while navigating a visual scene to locate linguistically specified targets. We investigate when and how listeners use object names in visual-search strategies across three visual world experiments, varying the presence and location of an added visual memory demand. The results suggest that as long as objects in the display can be visually inspected throughout the trial, participants do not linguistically encode those objects. We suggest that instead they use the visual environment as an external memory, mapping the spoken word onto potential referents and using perceptual visual routines automatically triggered by the spoken word. The results are discussed in terms of flexible and efficient allocation of memory resources in natural tasks that combine language and vision.

Keywords: visual search, spoken-word recognition, memory

Introduction

Flexible use of memory and attentional resources is a hallmark of performance in natural tasks. For example, in a classic set of studies, Ballard, Hayhoe and Pelz (1995) monitored eye movements in a block-copying task. Participants used Duplo™ blocks from a *resource area* to duplicate a pattern from a *model area* in an adjacent *workspace.* The only task constraint was that only one block could be moved at a time. The striking result was that participants typically fixated twice on each block in the model prior to a block move. The first fixation identified the color of the block, which was then picked up from the resource area. Position was encoded on a second fixation to the block in the model area. Ballard et al. argued that this pattern reflects a tradeoff between the resources needed to encode and maintain the model block in memory and those involved in making multiple eye movements. Using the display as an "external memory" was less resource-demanding than encoding and binding two dimensions (color and position) and holding them in working memory. When eye movements were made more "expensive" by increasing the distance between the workspace and model areas, participants made fewer fixations and relied on richer memory representations.

Interlocutors face similar resource allocation tradeoffs when they use language to converse about a co-present visual world. Consider for example a task in which participants follow verbal instructions to select a co-present referent, as in typical "visual world" studies of spoken-word recognition: Four or more pictures of objects with common names are shown in a display, as in Figure 1, and participants are instructed to click on one of the objects. This setup combines two tasks that could flexibly draw upon different resources: processing spoken language and searching a co-present visual display. Participants could linguistically encode the displayed pictures, allowing them to match the unfolding speech against these sound-based representations (Huettig & McQueen, 2007). Alternatively, participants could rely on perceptual/visual routines triggered by the unfolding speech (Dahan & Tanenhaus, 2005; Salverda, Brown & Tanenhaus, 2011) to guide eye movements to potential search targets in the scene. If these routines were automated (cf. Salverda & Altmann, 2011), participants would be able to use the co-present display as an external memory, thereby avoiding interference between processing the unfolding speech and maintaining a phonological representation of the picture names (cf. Brooks, 1968).

In order to detect the use of phonological encoding of objects in a visual display under different conditions, we selected pictures with two names that are judged to be appropriate (e.g., couch/sofa), while ensuring that one of the names is strongly preferred. Phonological encoding of such a picture would typically result in the retrieval of its dominant name but not its subordinate name (as in overt picture naming). In Experiment 1, the name of the target picture overlapped phonologically with either the dominant or subordinate name associated with the competitor (e.g., *cow* or *soda*, respectively). We observed clear competition effects, as indexed by looks to the competitor, with no difference in looks to the competitor as a function of whether the target overlapped with its dominant or subordinate name, suggesting that visual search was not mediated by encoded picture names.

Building on classic work on the role of linguistic (i.e., sound-based) coding for temporarily maintaining visual stimuli in memory (Posner & Mitchell, 1967; Posner, 1978; Conrad, 1964), in a second study we encouraged name encoding of one picture in the display by masking it before the spoken instruction began. When the synonym picture was masked (Experiment 2a), strong name-typicality effects emerged, suggesting that participants had retrieved the dominant name of the picture and used it to guide visual search. When an unrelated distractor was masked (Experiment 2b), we did not see typicality effects for the synonym picture. We discuss these results in terms of resource allocation in tasks combining language and vision.

Methods

Materials

Forty workers on Amazon Mechanical Turk provided a name for each of 45 images with multiple names. We then selected images for which the two most frequently provided names were chosen by at least 75% of the participants, and the ratio between the first and second most frequently chosen names was between .65 and .95. We use this ratio as a metric of name typicality, referring to the more frequently chosen name as the *dominant* name (e.g. couch), and the less frequently chosen name as the *subordinate* name (e.g. sofa). We refer to these items as *synonym competitors*[1].

We then collected picture-name agreement ratings for the entire set of pictures. Participants saw 45 image-name pairs. Items were counterbalanced so that each subject saw a particular image with either its dominant or subordinate name and saw an approximately equal number of dominant and subordinate picture-name pairs. Each trial started with the presentation of a written name for 1500 ms, followed by a button which, when clicked, displayed the associated image. Participants were then asked, "How good is this name for the object?", and they provided a rating on a scale ranging from 1 ("Very bad") to 7 ("Very good").

We selected a subset of 14 images for use as synonym competitors in our experiments. We chose pictures with average picture-name agreement ratings higher than 5 and the most similar ratings for the dominant and subordinate name. Mean ratings were 6.19 for dominant names and 5.90 for subordinate names.[2]

We then selected images representing a phonological cohort of each of the two names associated with a synonym competitor (e.g., *cow* for *couch* and *soda* for *sofa*). These pictures were search targets on critical trials in our experiments.

A native speaker of American English recorded two spoken instructions for each of the 14 critical trials and one instruction for 48 additional filler trials. The name of each target (e.g. *The cow*) was recorded separately and the audio was spliced in front of the sentence frame "is the target." There was some variation in the degree of phonological overlap between the names associated with a synonym object (e.g. *couch*/*sofa*) and their dominant and subordinate cohorts (*cow* and *soda*). In order to quantify the expected point of disambiguation between associated cohort pairs we conducted a gating study on Amazon Mechanical Turk. Forty subjects heard each of the target words in fragments of increasing duration. Each fragment started at the onset of the determiner, and subsequent fragments increased in duration by 40 ms. Following each presentation, the name of the target (e.g., *cow*) and the name of its associated synonym competitor (*couch*) appeared on the screen. Participants indicated which of the two words the fragment corresponded to and rated their confidence on a 9-point scale, with an option to select "Absolutely certain" which ended the trial. For each target word, we operationalized the point of disambiguation as the first gate at which participants provided a confidence rating of at least 7 for a correct response and continued to do so in subsequent fragments. The mean point of disambiguation across all subjects was 299 ms for the dominant targets, and 300 ms for the subordinate targets.

Experiment 1

Participants. Forty-eight students from the University of Rochester participated in each experiment, receiving $10 as compensation. Each participant was a native speaker of American English with normal or corrected-to-normal vision and normal hearing.

Design. In addition to our 14 critical trials we included 48 filler trials to balance the presentation of our stimuli. Sixteen filler trials included an object with two possible names from our norming study that had not been selected as a synonym competitor, and in half of those trials, this object was the target. This ensured that when a synonym competitor was present in a trial, it was roughly equally as likely to be the target (8 out of 30 (i.e., 16 fillers + 14 critical) trials) as any of other pictures. On half of those 16 fillers, a phonological competitor of one of the object's names was included in the display. On an additional 12 filler trials, a phonological cohort pair was presented along with two unrelated objects. This ensured that when the display included a cohort pair, a cohort was nearly as likely to be the target as the other two pictures (18 out of 34 (i.e., 12 + 8 + 14) trials). In the remaining 20 filler trials, all four items were unrelated. Target and visual mask locations in the filler trials were balanced so that the masked object was roughly equally likely to be the target than any of the other objects.

Each participant saw 14 critical trials, split between the dominant and subordinate condition. In 7 trials, the target (e.g. *cow*) phonologically overlapped with the dominant name associated with the synonym competitor (*couch*), and in 7 trials, the target (e.g. *soda*) overlapped with the subordinate name associated with the synonym competitor (*sofa*). Two counterbalanced trial lists varied, for each critical trial, which of the two possible targets was mentioned in the spoken instruction. The trial order and location of the objects in the visual display were randomized for each participant. All lists began with 5 non-critical trials to familiarize participants with the task and procedure.

[1] We note that not all of the items are actually synonyms.

[2] We were unable to find a set of items with perfectly balanced picture-name agreement ratings. A paired *t* test showed that there was a statistically significant difference across conditions in average picture-name agreement ratings ($t(13) = 3.43$, $p = 0.0022$). We note that the significantly higher ratings for the dominant names introduce a slight bias in favor of the dominant names. A bias in the other direction would have complicated interpretation of the absence of a name typicality effect: looks to the subordinate name might then be inflated because they were a better fit to the picture than the dominant name.

Procedure On each trial, participants saw a display with four pictures (see Figure 1a). After a short delay, a spoken instruction of the form *The X is the target* was presented over headphones. Eye movements were monitored using an EyeLink II head-mounted eye-tracking system, sampling at 250 Hz. A drift correction was performed every five trials. The trial ended when a participant clicked on one of the pictures in the visual display.

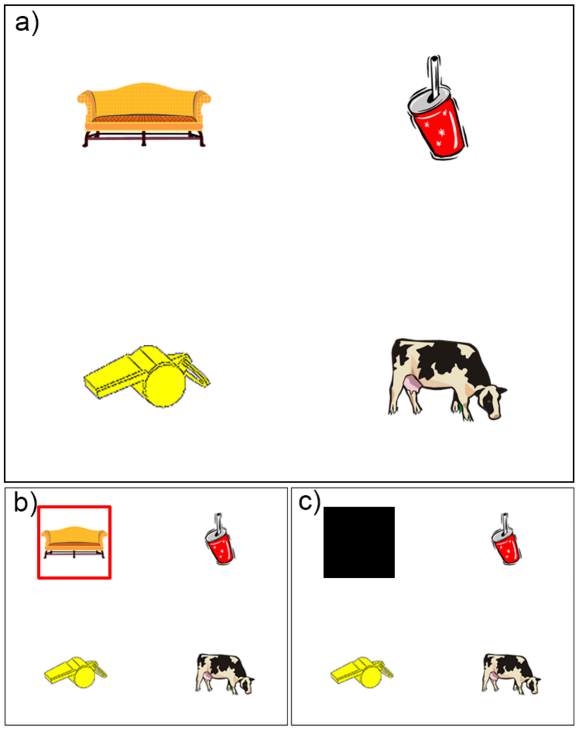

Figure 1. Panel A shows a schematic of a critical trial with the synonym picture (couch/sofa), dominant target (cow), subordinate target (soda) and unrelated distractor (whistle). In Experiment 1, pictures were displayed for 3 s before onset of the instruction, remaining on the screen throughout the trial. In contrast, in Experiment 2, one picture was masked: a red box appeared around the picture at trial onset (panel B), and after 1500 ms that picture was replaced by a mask (panel C). On critical trials in Experiment 2a, the synonym picture was highlighted and masked; on critical trials in Experiment 2b, the unrelated distractor was masked.

Results Trials with incorrect responses (1.3 % of the data) were excluded from analysis. Figure 2 presents the proportion of fixations over time to the target, synonym competitor and the distractors in the dominant and subordinate condition.

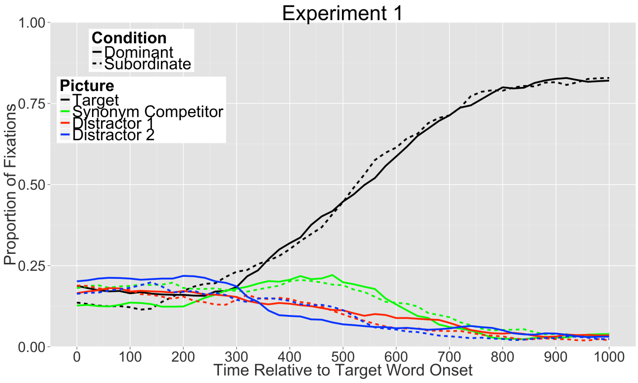

Figure 2. Proportion of fixations to the target, synonym competitor and distractors in Experiment 1.

In both conditions, shortly after 200 ms following the onset of the target word, fixations to the target and synonym competitor increased relative to those to the distractors. Target fixations continued to rise as the target word unfolded, while synonym competitor fixations merged with distractors around 700 ms after target word onset—roughly corresponding to 200 ms after the offset of the target word (whose average duration was 460 ms).

Analysis For each trial, we computed the mean proportion of fixations to the synonym competitor during a time window corresponding to the unfolding of the target word, offset by 200 ms (taking into account an estimate of the earliest linguistically-mediated saccades; see Salverda et al., 2014).

We analyzed the data using a linear mixed-effects regression model (LMEM) which predicted the empirical logit transform (Cox & Snell, 1970) of the ratio of proportions of fixations to the synonym competitor to the sum of the proportion of fixations to the synonym competitor and a distractor. This competitor preference ratio was taken to reflect the degree to which the synonym competitor was considered as a potential referent during the unfolding of the target word.

The model included fixed effects for name typicality (dominant vs. subordinate), the phonological overlap measure established by our gating study (in order to account for differences in phonological overlap across items), their interaction, and random intercepts by participant and item (i.e., synonym competitor) as well as random slopes for condition by participant and by item. This was the maximal random-effects structure motivated by our design which resulted in a converging model (Barr et. al, 2013). Model syntax was as follows, using the lmer function in the lme4 package (version 1.1-7; Bates et al., 2011) in R.

```
Competitor Preference ~ Typicality * Overlap
  + (1 + Typicality | Item)
  + (1 + Typicality | Subject)
```

Crucially, the regression model showed no significant effect of name typicality (β = 0.023, se = 0.131, $\chi^2(1) = 0.032$, $p = 0.858$)[3], phonological overlap (β = 0.010, se = 0.059, $\chi^2(1) = 0.0319$, $p = 0.858$), nor their interaction (β = 0.130, se = 0.119, $\chi^2(1) = 1.257$, $p = 0.262$).

Discussion The presence of competitor effects along with the absence of a name typicality effect is inconsistent with the idea that phonological representations of the pictures played a primary role in mapping spoken target words onto their referents. We suggest instead that participants used the visual display as an external memory and that perceptual representations or routines triggered by processing the unfolding word mediated identification of the referent (cf. Dahan & Tanenhaus, 2005; Salverda et al., 2011).

If generation of name-based codes is strategic, and if such phonological codes can be used to map a spoken target word onto a picture in a visual display, typicality effects should emerge when memory demands encourage linguistic coding of the synonym picture. Experiments 2a and 2b examined this hypothesis.

Experiment 2

We created memory demands by masking one of the pictures in the display, using a red box to signal which picture would be masked upon display onset but prior to the presentation of the spoken instruction (see Figure 1). In addition, there was a time limit for each trial. Two seconds after the offset of the target word, the experiment automatically advanced to the next trial. Participants were therefore instructed to respond quickly to the spoken instruction. We hypothesized that participants would linguistically encode (name) the to-be-masked picture to maintain it in memory, and that this would typically result in the retrieval of its dominant name.

In Experiment 2a, the synonym competitor picture was masked on critical trials. We predicted that this would result in more looks to the synonym object in the dominant condition, where the name of the target overlapped with the most accessible name associated with the synonym competitor, compared to the subordinate condition, where the name of the target overlapped with the synonym competitor's less accessible name. Thus, a typicality effect should emerge under these conditions.

In Experiment 2b we masked an unrelated picture, creating the same processing and memory demands in the task while leaving the synonym picture in the visual display. We predicted that we should again see the same pattern of results as in Experiment 1, that is, no difference between competitor effects in the dominant and subordinate conditions. This result would argue against an alternative explanation that potential typicality effects in Experiment 2a would be due to increased task difficulty and memory load associated with the presence of the visual mask rather than specific effects of linguistic encoding of the synonym picture.

Procedure The auditory and visual stimuli were identical to Experiment 1. Aside from the addition of the time limit (each trial automatically ended two seconds after word offset), the timing of the trials was also the same, with the onset of the instruction occurring three seconds after display onset. On each trial, one of the pictures was marked by a red outline at the onset of the visual display. Participants were instructed that marked pictures would be masked. After 1500 ms, a visual mask appeared on top of the object. On half of the trials, the mask was fully opaque. On the other half, it was partially transparent, making the masked object difficult to identify. This introduced a visual memory demand, intended to encourage a name encoding strategy (cf. Posner et al., 1996). On critical trials, the mask was always fully opaque. Filler trials were constructed so that the masked picture was the target on 16 trials and the location of the mask did not predict the location of the target.

Experiment 2a

Results Trials with incorrect responses (0.7% of the data) were excluded from analysis. Figure 3 plots the proportion of fixations over time to the target, the synonym competitor and the unrelated distractors. In the dominant condition, fixations to the synonym competitor increased around 200 ms and merged with distractor fixations at around 600 ms. In contrast, in the subordinate condition, fixations to the synonym competitor did not increase substantially during the processing of the target word.

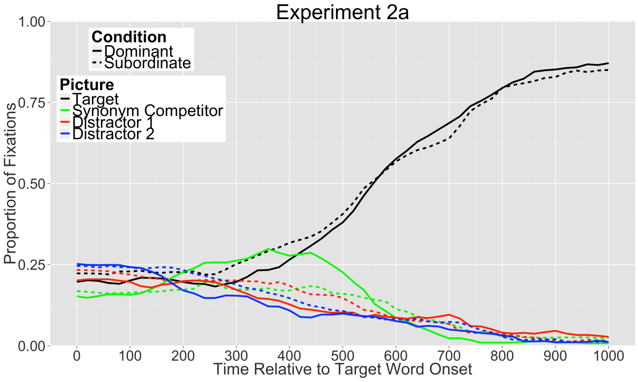

Figure 3. Proportion of fixations to the target, synonym competitor and distractors in Experiment 2a.

Analysis We conducted analyses that parallel those conducted for Experiment 1 using the time window between 200 ms after word onset to 200 ms after word offset. A LMEM predicted the same outcome measure and used the same fixed effects and random-effects structure as that used to analyze the data of Experiment 1.

[3] P-values were computed by performing a likelihood ratio test, using R's ANOVA function, between the model with and without the particular fixed effect being ascertained.

```
Competitor Preference ~ Typicality * Overlap
  + (1 + Typicality | Item)
  + (1 + Typicality | Subject)
```

There was a significant main effect of name typicality (β = -0.309, se = 0.148, $\chi^2(1) = 4.282$, $p = 0.038$), showing that participants were more likely to fixate the synonym competitor relative to a distractor in the dominant condition than in the subordinate condition. There was neither a significant effect of phonological overlap (β = 0.046, se = 0.066, $\chi^2(1) = 0.562$, $p = 0.453$) nor a significant interaction between typicality and overlap (β = 0.071, se = 0.132, $\chi^2(1) = 0.322$, $p = 0.57$).

Experiment 2b

In Experiment 2b, an unrelated distractor object was masked on critical trials. We constructed filler trials so that throughout the experiment, it was not predictable whether a synonym object would be masked. The same logic from Experiment 2a predicts that participants should phonologically encode the distractor object. Importantly, we predicted that this encoding should not affect fixations to the synonym competitor, since this object remains on the screen throughout the trial. If participants do not linguistically encode the synonym competitor, we expect that there will be no difference in fixations to that competitor as a function of name typicality—i.e., whether the target overlaps with the synonym competitor's dominant or subordinate name.

Results Trials with incorrect responses (2.1% of the data) were excluded from the analysis. Figure 4 plots the proportion of fixations over time to the target, synonym competitor and the unrelated distractors. Fixations to the synonym competitor showed a similar pattern to that observed in Experiment 1. Between 200 and 700 ms after word onset, there were more fixations to the competitor relative to the unmasked distractor in both the dominant and subordinate condition. Fixations to the masked distractor dropped sharply from 200 ms after word onset.

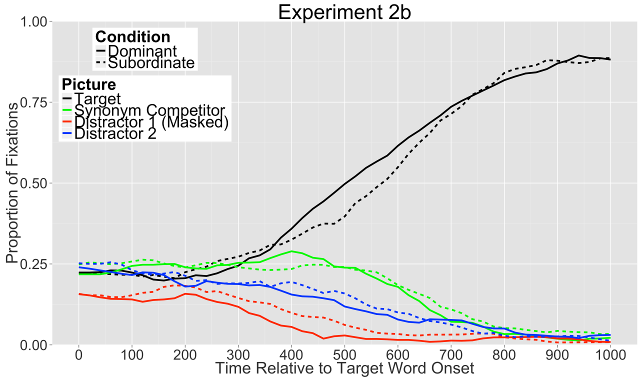

Figure 4. Proportion of fixations to the target, synonym competitor and distractors in Experiment 2b.

Analysis As in the previous experiments, our analysis was conducted on the time window from 200 ms after word onset to 200 ms after word offset. A LMEM predicted synonym competitor preference using name typicality (dominant vs. subordinate), phonological overlap and their interaction as predictors, and random intercepts by participant and item (the maximal converging random effects structure).

```
Competitor Preference ~ Typicality * Overlap
  + (1 | Item)
  + (1 | Subject)
```

There was no significant effect of typicality (β = -0.043, se = 0.096, $\chi^2(1) = 0.201$, $p = 0.654$), suggesting that participants fixated the synonym competitor as often in the subordinate condition as they did in the dominant condition. There was also no significant effect of phonological overlap (β = -0.025, se = 0.056, $\chi^2(1) = 0.178$, $p = 0.673$) or the interaction between typicality and overlap (β = 0.033, se = 0.110, $\chi^2(1) = 0.067$, $p = 0.796$).

Discussion

In Experiment 2a, the masked synonym competitor was a strong competitor when the target word was associated with its dominant name, but not when the target word was associated with its subordinate name. This suggests that because participants knew that the picture would not be available for use as an external memory source, they linguistically encoded it. In Experiment 2b, overall memory demands due to masking were similar but it was a distractor that needed to be maintained in memory. Under these conditions, we replicated the pattern observed in Experiment 1, namely, equivalent competition effects for dominant and subordinate synonym competitors.

General Discussion

We demonstrated that name typicality does not affect reference resolution unless participants encode the name of the picture in order to maintain it in memory. This suggests that the mapping of a spoken word onto a visually co-present referent was mediated by visual representations or routines, triggered as the spoken word unfolded, rather than pre-encoded object names. We argue that this result can be situated within the broader issue of resource allocation in natural tasks, specifically those involving language and vision.

We noted that linguistically encoding a co-present visual display might interfere with processing the spoken language. In contrast, activating visual routines would draw attention to relevant objects when interlocutors are talking about the co-present world. Much of joint language behavior, however, is not about co-present entities and events. What role then might visual routines play? We hypothesize that automatic visual/perceptual routines form a substrate for perceptually-based (non-linguistic) internal models that support comprehension without interfering with processing the ongoing language.

Similarly, a number of theorists have suggested that there might be a close link between production and comprehension, with the production system generating predictions about likely upcoming signals (Dell & Chang, 2013; Pickering & Garrod, 2013). If that is the case, then the listener's production system would be focused on generating expectations about what the speaker might say. Given the uncertainty about what the speaker might say in the current situation (e.g., any the four pictures could be the target, and pictures in general can often be referred to by different names), making specific predictions might be inefficient. However, if the speaker had already used a particular name to refer to a picture, she would likely use the same name to refer to it again. Under these circumstances, and if repeated reference to a picture is likely, encoding the name of that picture may confer a processing advantage—a hypotheses we are exploring in ongoing research.

By comparison, in Experiment 2a, looks to the dominant synonym competitor rise quickly at the onset of the spoken word and return to baseline quickly compared to Experiment 1. This suggests that having a phonological code could facilitate not only linking predictable words to visual referents, but also efficiently removing hypothesized referents from the search space following a signal-prediction mismatch. Indeed, we see hints of such an effect in a rapid drop of looks to the masked (unrelated) picture in Experiment 2b.

We also note that in an interactive conversation, interlocutors are simultaneously both speakers and listeners. If we view the production system as a limited, shared resource, we might expect a tradeoff between generating a name for a visual referent and predicting the other interlocutor's utterances. If both interlocutors are likely to refer to the same referents, efficiently aligned referential expressions would leave the production system available for generating predictions for comprehension. This offers a slightly different perspective on the importance of negotiating common referring expressions (conceptual pacts) under conditions of uncertainty. While these ideas are speculative, we believe that extensions of the present paradigm will provide a method for empirically examining questions about resource allocation in situations involving combined use of language and vision.

Acknowledgments

Thanks to Elika Bergelson, John Trueswell and Delphine Dahan for helpful discussions. Partially supported by NIH grant HD073890 to MKT.

References

Ballard, D., Hayhoe, M., & Pelz, J. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience 7*(1), 66-80.

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*(3), 255-278.

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

Brooks, L. R. (1968). Spatial and verbal components of the act of recall. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *22*(5), 349-368.

Conrad, R. (1964). Acoustic confusions in immediate memory. *British Journal of Psychology*, *55*(1), 75-84.

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data*: *Second edition.* Boca Raton: CRC Press LLC.

Dahan, D., & Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic Bulletin & Review*, *12*(3), 453-459.

Dell, G. S., & Chang, F. (2014). The P-chain: relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120394.

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language*, *57*(4), 460-482.

Pickering, M. J., & Garrod, S. (2013). Forward models and their implications for production, comprehension, and dialogue. *Behavioral and Brain Sciences*, *36*(4), 377-392.

Posner, M. I. (1978). *Chronometric explorations of mind*. Oxford: Oxford University Press.

Posner, M. I., & Mitchell, R. E. (1967). Chronometric analysis of classification. *Psychological Review*, *74*(5), 392-409.

Salverda, A. P., & Altmann, G. T. M. (2011). Attentional capture of objects referred to by spoken language. *Journal of Experimental Psychology: Human Perception and Performance, 37(4)*, 1122-1133.

Salverda, A. P., Brown, M., & Tanenhaus, M. K. (2011). A goal-based perspective on eye movements in visual-world studies. *Acta Psychologica, 137*(2), 172-180.

Salverda, A. P., Kleinschmidt, D., & Tanenhaus, M. K. (2014). Immediate effects of anticipatory coarticulation in spoken-word recognition. *Journal of Memory and Language*, *71*(1), 145-163.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632-1634.