Immediate Integration of Syntactic and Referential Constraints on Spoken Word Recognition

James S. Magnuson (magnuson@psych.columbia.edu)
Department of Psychology, Columbia University
1190 Amsterdam Ave., MC 5501
New York, NY 10027 USA

Michael K. Tanenhaus (mtan@bcs.rochester.edu) and **Richard N. Aslin** (aslin@cvs.rochester.edu)
Department of Brain & Cognitive Sciences, University of Rochester
Rochester, NY 14627 USA

Abstract

We tested the hypothesis that syntactic constraints on spoken word recognition are integrated immediately when they are highly predictive. We used an artificial lexicon paradigm to create a lexicon of nouns (referring to shapes) and adjectives (referring to textures). Each word had phonological competitors in both form classes. We created strong form class expectations by using visual displays that either required adjective use or made adjectives infelicitous. We found evidence for immediate integration of form class expectations based on the pragmatic visual cues: similar-sounding words competed when they were from the same form class, but not when they were from different form classes.

Top-down constraints on word recognition

It is clear that we integrate top-down information when we interpret language. If someone tells us they put money in a bank, we understand that their money is in a vault and not buried next to a river. What is less clear is *when* and *how* we integrate top-down knowledge with bottom-up linguistic input.

One possibility is that language is processed in stages, with top-down information integrated after an encapsulated first-pass on the bottom-up input (e.g., Frazier & Clifton, 1996; Norris, McQueen & Cutler, 2000). The theory behind this genre of model is that optimal efficiency can be achieved by applying automatic processes that will almost always yield a correct result. In the rare event that the automatic result cannot be reconciled with top-down information, reanalysis would be required.

A second possibility is that top-down constraints are integrated immediately, with weights proportional to their predictive power (e.g., McClelland & Elman, 1986; MacDonald, Pearlmutter & Seidenberg, 1994; Tanenhaus & Trueswell, 1994). The theory behind constraint-based approaches is that a system can be made more efficient by allowing any sufficiently predictive information source to be integrated with processing as soon as it is relevant.

While a variety of results support constraint-based theories of sentence processing (see MacDonald et al., 1994), there is reason to believe that spoken word recognition is initially encapsulated from top-down constraints. Swinney (1979) and Tanenhaus, Leiman & Seidenberg (1979) provided the seminal results on this issue by examining whether all homophones are activated independent of context. Tanenhaus et al. presented participants with spoken sentences that ended with a syntactically ambiguous word (e.g., "they all rose" vs. "they bought a rose"). If participants were asked to name a visual target immediately at the offset of the ambiguous word, priming was found for associates both of the alternative suggested by the context (e.g., "stood" given "they all rose") and of homophones that would not fit the syntactic frame (e.g., "flower"). Given a 200-ms delay prior to the presentation of the visual stimulus, priming was found only for associates of the syntactically appropriate word. This suggests that lexical activation is initially based only on bottom-up information, and top-down information is a relatively late-acting constraint.

Tanenhaus & Lucas (1987) argued that this made sense given the predictive power of a form-class expectation. Knowing that the next word will be one of tens of thousands of nouns would afford virtually no advantage for most nouns (those without homophones in different form classes). Furthermore, expectations for classes like noun or verb might be very weak because modifiers can almost always be inserted before either class (e.g., "they just rose", "they bought a very pretty red rose"; cf. Shillcock & Bard, 1993).

Shillcock & Bard (1993) pointed out that there are form classes that should be more predictive than *noun* or *verb*, because they have few members: those made up of closed-class words. They examined whether /wʊd/ in a sentence context favoring the closed-class item, "would" (e.g., "John said that he didn't want to do the job, but his brother would, as we later found out"), would prime associates of its homophone, "wood", such as "timber" (compared with a context like "John said he didn't want to do the job with his brother's

wood, as we later found out"). They found priming for "timber" given the open-class context (favoring "wood") immediately after the offset of /wʊd/, but not given the closed-class context. The same result held when they probed half-way through the pronunciation of /wʊd/. This suggests the closed-class context was sufficiently constraining to bias the earliest moments of word recognition. A cloze test (in which participants were asked to supply the next word given the sentence contexts up to the word just prior to "would" or "wood", with the understanding that the word they supplied would not be the last in the sentence) confirmed that the closed-class context was much more predictive. Participants provided words of the same form class as the target most of the time for both cases, but were much more likely to provide the target given the closed-class context than the open-class context.

Shillcock & Bard's result is consistent with the constraint-based view that top-down information sources are integrated early in processing when they are sufficiently predictive. In the current experiment, we tested the hypothesis that even form class expectations for open-class words could constrain word recognition given a context with sufficient predictive power.

The Experiment

We hypothesized that form class could be sufficiently predictive to constrain initial activation if it were combined with strong visual and pragmatic expectations. For example, if there are four objects on a table – a brown purse, a purple book, a red ashtray, and a blue pen – and we ask you to pick one up, you would have strong expectations about how specific we would be in making reference to an item. For example, if we wanted the purse, you would expect to be asked, "pick up the purse" rather than "pick up the brown purse." Because of such conversational pragmatics (Grice, 1975), we would not expect subjects to experience strong competition between "purple" and "purse" as they hear "pur—," since if we wanted the book, we would ask for "the book," not "the purple book." But if there were brown and red purses, and purple and green books, given "pick up the pur–" we would expect little competition from *purse* – subjects would have a strong expectation to hear an adjective in this case.

Constructing such an experiment with real words poses significant problems. While there are many examples of cross-form class competitors in English, there are relatively few that are highly imageable and thus appropriate for our pragmatic manipulation. Even among these few, there is high variability in factors such as frequency and word length (e.g., purple-purse, dotted-dog, tan-tambourine, rough-rum).

Therefore, we extended an artificial lexicon paradigm that we previously developed to study the lexical neighborhoods of spoken words (Magnuson, Dahan, Allopenna, Tanenhaus & Aslin, 1998). An artificial lexicon allows precise control over such dimensions as phonological similarity and frequency of occurrence, as well as visual aspects of stimuli.

Table 1: The artificial lexicon.

	NOUN (shape)	ADJ (texture)	
1	pibo	pibʌ	1
2	pibe		
3	bupo	bupʌ	2
		bupe	3
4	tedu	tedi	4
		tedɛ	5
5	dotɛ	doti	6
6	dotu		
7	kagæ	kaga^{i}	7
		kagu	8
8	ga^{w}ku	ga^{w}kæ	9
9	ga^{w}ka^{i}		

We created a lexicon of nouns (referring to novel shapes) and adjectives (referring to textures). The lexicon (shown in Table 1) contained phonemic cohorts (e.g., /pibo/ and /pibʌ/) in different syntactic categories (e.g., /pibo/ was a noun and /pibʌ/ was an adjective) or the same category (e.g., another noun was /pibe/). Thus, the artificial lexicon allowed us to compare phonological competitors in same or different form classes with similarity precisely controlled. (Note that there are even fewer examples of real words with comparable phonological competitors in the same form class and another, and the possible sets are quite heterogeneous, e.g.: purple-purse-person, tattered-tan-tambourine.)

Participants learned the lexicon over two days of training. Instructions were given in an English context, with English word order (e.g., "click on the /pibʌ/ [adj] /tedu/ [noun]"). We created conditions in which the visual context provided strong syntactic expectations by constructing contexts in which adjectives were required (e.g., two examples of the shape associated with /pibo/, but with two different textures) or infelicitous (e.g., two different shapes, making the adjective superfluous, even if the shapes have different textures). If syntactic expectations in conjunction with pragmatic constraints embodied in the visual display can constrain word recognition early in processing, we should observe competition effects only between cohorts from the same syntactic form class.

Methods

Participants

Eight native speakers of English who reported normal or corrected-to-normal vision and normal hearing were paid for their participation. Participants attended sessions on two consecutive days.

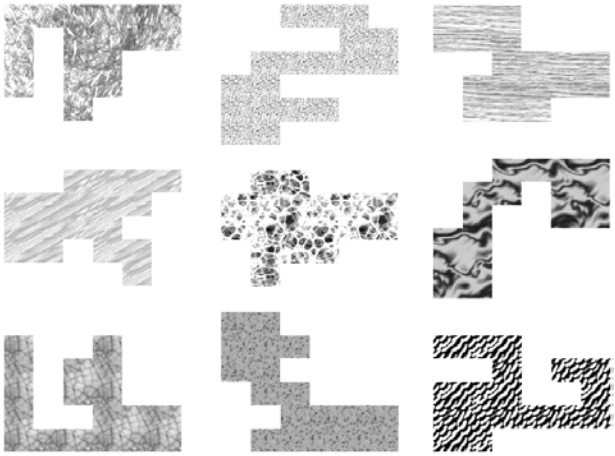

Figure 1: The 9 shapes and 9 textures.

Materials

The linguistic materials consisted of the 18 artificial words (9 nouns, referring to shapes, and 9 adjectives referring to textures) shown in Table 1. The auditory stimuli were produced by a male native speaker of English in a sentence context (e.g., "Click on the /bupe tedu/."). The stimuli were recorded using a Kay Lab CSL 4000 with 16 bit resolution and a sampling rate of 22.025 kHz. The mean duration of the "Click on the…" portion of the instruction was 475 ms for adjective instructions, and 402 ms for noun instructions. For adjective instructions, mean adjective duration was 487 ms, and mean noun duration was 682 ms. For noun instructions, mean noun duration was 558 ms.

We examined the neighborhoods our artificial words would fall into were they real words of English; none would be in a dense English neighborhood (9 had 0 neighbors, and 7 had 1 neighbor). (See Magnuson [2001] for evidence that artificial and native lexicons do not interact, even when artificial items are constructed to be maximally similar to real words.) The visual materials consisted of unfamiliar shapes generated by randomly filling 18 contiguous cells in a 6x6 grid. We selected a set of 9 subjectively dissimilar shapes. These shapes provided referents for the nouns. In addition, 9 textures were selected from among the set distributed with Microsoft PhotoDraw. Figure 1 shows each of the 9 shapes, with a different one of the 9 textures applied to each (note that picture quality was substantially higher on the computer display). Names were randomly mapped to shapes and textures for each participant.

Eye tracking

During the tests (see Procedure), eye movements were monitored using a SensoriMotoric Instruments (SMI) EyeLink eye tracker, which provided a record of point-of-gaze in screen coordinates at a sampling rate of 250 hz. Saccades and fixations were coded from the point-of-gaze data using SMI's software.

Eye movements were used because they are closely time-locked to speech in a properly constrained task. Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy (1995) found time locked fixations when subjects followed spoken instructions to perform a visually-guided task (e.g., "pick up the candle"). Because the subject must foveate the target item in order to efficiently follow the instruction, there is a functional link between the speech stimulus and dependent measure. This link avoids the pitfalls of interpreting eye movements described by Viviani (1990).

Allopenna, Magnuson & Tanenhaus (1998) extended this work to a time-course issue in spoken word recognition. Whereas studies using more conventional tasks had failed to find evidence for the activation of rhymes during lexical competition, eye tracking proved sensitive enough to detect the robust (if relatively weak) rhyme activation predicted by various models (e.g., Luce & Pisoni, 1998; McClelland & Elman, 1986). Dahan, Magnuson & Tanenhaus (2001) applied the approach to a debate regarding frequency effects in spoken word recognition. Competing theories made conflicting predictions at the level of time course; for example, some argued it kicked in as a late bias (Connine, Blasko & Titone, 1993). Dahan et al.'s eye tracking measures demonstrated that frequency has a continuous but gradual influence from the earliest moments of processing, leading to the appearance of a late locus in less sensitive paradigms.

The eye tracking paradigm imposes different constraints than more conventional paradigms, such as lexical decision. In a conventional task, the stimuli are typically decontextualized; there is nothing about the task that predicts what word one might hear next. In the eye tracking paradigm, the stimuli are presented in the context of a display of items. While this allows more naturalistic tasks, it might also allow strategic processing. For example, participants might activate lexical representations in response to the visual display prior to any bottom up information, or the displayed set of items might provide a verification set to guide recognition. There is no evidence for lexical activation prior to the bottom-up signal; fixation proportions map precisely onto emerging phonetic similarity over time. We have also found that recognition in this paradigm is not based on lexical activations constrained to the displayed items: artificial lexical items (Magnuson, Tanenhaus, Aslin & Dahan, 1999, in preparation) and real words (Magnuson, 2001) in dense neighborhoods (i.e., with many or very frequent neighbors) are recognized more slowly than words from sparser neighborhoods, even when the neighbors are not displayed. This suggests the representations of the neighbors were activated and competed for recognition.

In summary, eye movements provide an extremely sensitive time course measure of lexical activation and competition. We need just such a measure to resolve the time course debate we are concerned with here: when

are top-down and bottom-up information integrated during spoken language understanding?

Procedure

Participants were trained and tested in sessions on two consecutive days. Each session lasted between 90 and 120 minutes. On day 1, participants were trained first on the nouns in a two-alternative forced choice (2AFC) task. On each trial, two shapes appeared (both with solid black texture) and the participant heard an instruction to click on one (e.g., "click on the bupo"). The auditory stimuli were presented binaurally through headphones (Sennheiser HD-570) using standard Macintosh Power PC digital-to-analog devices.

When the subject clicked on an item, one item disappeared, leaving the correct one, and its name was repeated. There were 14 repetitions of each item, split into 3 blocks of 48 trials. Items were not repeated on consecutive trials, and were ordered such that every item was repeated 7 times every 72 trials. Following the 2AFC blocks, noun training continued with 3 blocks of 4AFC, with identical ordering constraints and numbers of trials. Each shape appeared equally often as a distractor.

Adjective training then began. First, participants saw two exemplars of one shape, with different textures. They heard an instruction, such as "click on the bupe pibo". Since they already knew that, e.g., "pibo" referred to one of the shapes, participants found it transparent that "bupe" referred to one of the textures. As in the noun training, after they clicked on one item, the incorrect one disappeared and the full name was repeated. Each adjective and each noun was a target in 8 trials in each block; each adjective was randomly paired with 8 different nouns in each block. After three 48-trial 2AFC blocks, there were three 4AFC blocks, with four exemplars of the same shape with four different textures. These were followed by three more blocks of 4AFC, but with two exemplars each of two shapes, each with a different texture (requiring participants to recognize both the adjective and noun).

After this, a more complex training regime began. On some trials, four different shapes appeared. On others, two pairs of shapes appeared. On every trial, each shape had a different texture. On trials with two pairs of shapes, an adjective was required to make unambiguous reference, and the full referent was specified on such trials (e.g., "click on the bupe pibo"). On trials with four different shapes, the adjective was not required – each item could be identified unambiguously by the name of the shape, and so only the noun was specified in the instruction (e.g., "click on the pibo"). Using the adjective would be infelicitous, on Grice's (1975) maxim of quantity (one should not over-specify, which is the observed tendency in natural conversation). Each adjective was repeated 8 times in every block of 144 trials, paired each time with a different, randomly selected noun. Each noun was repeated as the target item 8 times in the 4-noun trials. Trials were presented in blocks of 48. Participants completed 3 blocks of this mixed training on Day 1. On Day 2, they completed 12 more, which comprised the entire training phase on Day 2.

After each 48-trial block, the participant saw a summary of his or her accuracy in that block. To motivate participants, we told them that each training segment would continue until they reached 100% accuracy. Typically, we moved to each successive training phase after the number of blocks listed above for each segment, except in a few rare cases where participants were below 90% accuracy after the specified number of blocks, in which case training continued for another 1-2 blocks.

Each day ended with a 4AFC test with no feedback. We tracked participants' eye movements during the test. There were six basic conditions in the test. In the *noun baseline condition*, there were four different shapes, and no shape's or texture's name was a competitor of the target noun. In the *noun plus noun cohort condition*, there were four shapes, and one of them was a cohort to the target (e.g., the target might be /pibo/, and /pibe/ would also be displayed), but no shape had the target's adjective cohort texture applied (e.g., no shape would have the /pibʌ/ texture). In the *noun plus adjective cohort condition*, four different shapes were displayed. The noun cohort was not displayed, but the adjective cohort was (e.g., a distractor might be /pibʌ tedu/). In these conditions, the instruction would only refer to the noun (e.g., "click on the pibo").

In the other three conditions, two exemplars of two different shapes were displayed, requiring the adjective to be used in the instruction. In the *adjective baseline condition*, none of the distractor textures were cohorts of the target, and neither were any of the nouns. In the *adjective plus adjective cohort condition*, one of the non-target textures was a cohort to the target (e.g., the target might be /tedi dotu/, and one non-target might be /tedɛ bupo/), but no noun cohorts of the target would be displayed. In the *adjective plus noun cohort condition*, none of the distractors would have textures that were cohorts to the target texture, but a noun cohort would be displayed (e.g., given /tedi dotu/ as the target, /bupe tedu/ might be included).

The following scheme was used to ensure that each adjective and target appeared 9 times as targets in the test. Note that nouns and adjectives either had one competitor in each form class, or two in the opposite form class. Nouns with noun cohorts appeared in six *noun baseline* trials, two *noun plus noun cohort* trials, and once in the *noun plus adjective cohort condition*. Nouns with two adjective cohorts appeared in 7 *noun baseline* trials, 0 *noun cohort* trials, and two *noun plus adjective cohort* trials. The same pattern was used with adjective conditions, giving a total of 162 test trials.

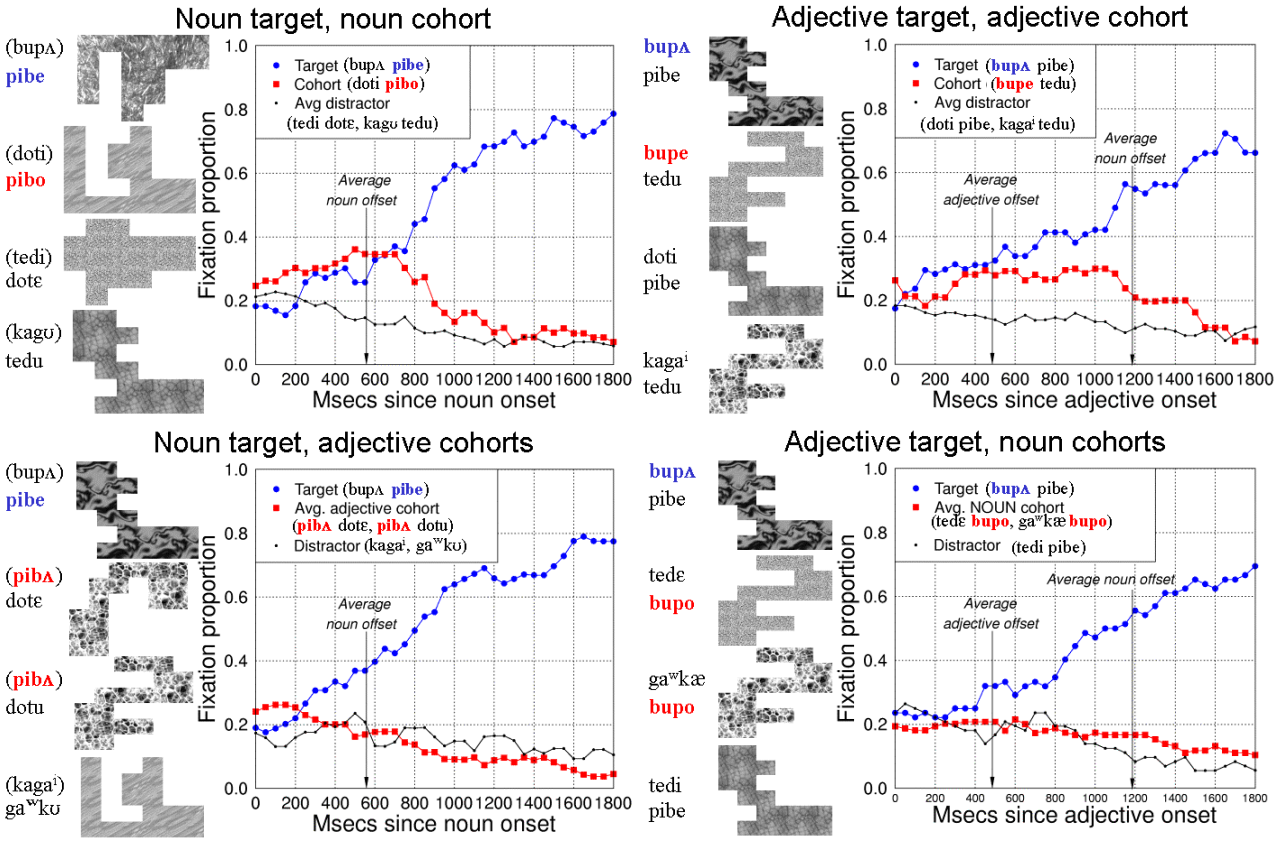

Figure 2: Results from the 4 critical conditions. The top panels show competition between within-class cohorts. The bottom panels show the failure to find competition for cohorts from different form classes.

Results

Participants attained high accuracy quickly (two failed to reach ceiling levels of accuracy, performing at less than 90% correct on the test on Day 2, and their data was excluded from the analyses). Mean accuracy on nouns and adjectives was 96% at the end of Day 1, and 98% at the end of Day 2. The results from the test on Day 2 are shown in Figure 2. Examples of possible stimulus items are shown to the left of each panel (these would be arranged around the central fixation cross in an actual experimental display). Note that in the cross-form class conditions (*noun with adjective cohorts* and *adjective with noun cohorts*) there were two cohorts in the display. This was necessary in the case of the *adjective plus noun cohort condition*; in order for the display to demand that an adjective be used, two exemplars of two different shapes had to be displayed. To make the *noun plus adjective cohort condition* comparable, two items were displayed with textures whose names were cohorts to the noun target.

The results show strong, immediate effects of the form-class constraints on lexical access. Compare the upper and lower panels of Figure 2. While strong cohort effects are apparent in the upper panels (the within-form class competitor conditions), there is no evidence for cohort effects in the lower panels (between-form class conditions). Analyses of variance on mean fixation proportion in the noun conditions over the window from 200 ms (where we first expect to see signal-driven fixations, since it takes 150 – 180 ms to plan and launch saccades in much simpler tasks) to 1400 ms (where the target proportions asymptote) confirm the trends. There was a reliably greater proportion of fixations to the cohort than to the distractors in the *noun plus noun cohort condition* (cohort=.25, mean distractor=.12; $F(1, 11)=10.16$, $p=.009$), but not in the *noun plus adjective cohort condition* (cohort=.15, mean distractor=.15). The same was true for the adjective conditions, over the window from 200 to 1800 (the window was extended because of the longer lag prior to disambiguation). There were reliably more fixations to the cohort in the *adjective plus adjective cohort condition* (cohort=.22, mean distractor=.15; $F(1,11)=7.2$, $p=.02$), but not in the *adjective plus noun cohort condition* (cohort=.16, mean distractor=.15, $p=.59$).

Discussion

The results demonstrate that higher-level linguistic constraints (in this case, syntactic expectations based on a visually-defined referential context) influence even the earliest moments of lexical access when the constraints are highly-predictive. Phonemically similar

items competed only when they were from the same form class. This suggests, contra strong modularity (e.g., Fodor, 1983), that lexical activation can be constrained given a highly informative context.

In future research, it will be important to establish the limits of such effects. It may be the case that form class constraints would be weaker were there more members of each form class (as predicted by the argument that a noun or verb expectation in English is an extremely weak constraint). We are currently exploring this possibility with an expanded lexicon.

It is possible that visual/pragmatic constraints swamp lexical activation, and turn the display into a verification set. To eliminate this possibility, we will use a neighborhood density manipulation. We should find faster increases in target fixations for items in sparse neighborhoods *in addition* to the form class/pragmatic effects observed here.

The timing of these sorts of effects will be informative about how different classes of constraints are integrated in real-time spoken word recognition. The current results provide a starting point for further explorations while demonstrating that the artificial lexicon paradigm can be adapted to a wide range of microstructural issues in spoken word recognition. Moreover, they suggest that the failure to find immediate effects in earlier studies does not reflect an architectural property of the word recognition system (i.e., encapsulation), but rather reflects the pattern predicted by constraint-based models when contextual constraints are only weakly predictive.

References

Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. Journal of Memory and Language, 38, 419-439.

Connine, C.M., Titone, D., & Wang, J. (1993). Auditory word recognition: Extrinsic and intrinsic effects of word frequency. Journal of Experimental Psychology: Learning, Memory & Cognition, 19, 81-94.

Dahan, D., Magnuson, J. S., and Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. Cognitive Psychology, 42, 317-367.

Frazier, L. & Clifton, C. (1996). Construal. Cambridge, MA: MIT Press.

Fodor, J. A. (1983). The Modularity of Mind. Cambridge, MA: MIT Press.

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.), Syntax and Semantics, Vol. 3, Speech Acts (pp. 41-58). NY: Academic Press.

MacDonald, M. C, Pearlmutter, N. J., and Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. Psychological Review, 101, 676-703.

Magnuson, J. S. (2001). The Microstructure of Spoken Word Recognition. Unpublished doctoral thesis, University of Rochester Department of Brain and Cognitive Sciences.

Magnuson, J. S., Dahan, D., Allopenna, P. D., Tanenhaus, M. K., and Aslin, R. N. (1998). Using an artificial lexicon and eye movements to examine the development and microstructure of lexical dynamics. In Gernsbacher, M.A., & Derry, S.J. (Eds.), Proc. of the Twentieth Annual Conference of the Cognitive Science Society, 651-656. Mahwah, NJ: Erlbaum.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., and Dahan, D. (1999). Spoken word recognition in the visual world paradigm reflects the structure of the entire lexicon. In M. Hahn & S. Stoness (Eds.), Proceedings of the Twenty First Annual Conference of the Cognitive Science Society, pp. 331-336. Mahwah, NJ: Erlbaum.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., and Dahan, D. (in preparation). The microstructure of spoken word recognition: Insights from investigations with artificial lexicons.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. Behavioural and Brain Sciences, 23, 299-370.

Saslow, M. G. (1967). Latency for saccadic eye movement. Journal of the Optical Society of America, 57 (8), 1030-1033.

Shillcock, R. C. and Bard, E. G. (1993). Modularity and the processing of closed-class words. In G. T. M. Altmann and R. Shillcock (Eds.), Cognitive Models of Speech Processing: The Second Sperlonga Meeting, pp. 163-185. Erlbaum.

Swinney, D. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. J. Verbal Learning & Verbal Behavior, 15, 545-569.

Tanenhaus, M. K., Leiman, J. M., and Seidenberg, M. S. (1979). Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. J. Verbal Learning & Verbal Behavior, 18, 427-441.

Tanenhaus, M. K., and Lucas, M. M. (1987). Context effects in lexical processing. Cognition, 25, 189-234.

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

Trueswell, J. C. & Tanenhaus, M. K. (1994). Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. In Clifton, C., Frazier, L. and Rayner, K. (Eds.) Perspectives in Sentence Processing. Erlbaum: Hillsdale, NJ.

Viviani, P. (1990). Eye movements in visual search: Cognitive, perceptual, and motor control aspects. In E. Kowler (Ed.), Eye Movements and Their Role in Visual and Cognitive Processes. Reviews of Oculomotor Research V4. Amsterdam: Elsevier.