Spoken Word Recognition in the Visual World Paradigm Reflects the Structure of the Entire Lexicon

James S. Magnuson (magnuson@bcs.rochester.edu)
Michael K. Tanenhaus (mtan@bcs.rochester.edu)
Richard N. Aslin (aslin@cvs.rochester.edu)
Delphine Dahan (dahan@bcs.rochester.edu)
Brain and Cognitive Sciences, University of Rochester, Meliora Hall, Rochester, NY 14627 USA

Abstract

When subjects are asked to move items in a visual display in response to spoken instructions, their eye movements are closely time-locked to the unfolding speech signal. A recently developed eye-tracking method, the "visual world paradigm", exploits this phenomenon to provide a sensitive, continuous measure of ambiguity resolution in language processing phenomena, including competition effects in spoken word recognition (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). With this method, competition is typically measured between names of objects which are simultaneously displayed in front of the subject. This means that fixation probabilities may not reflect competition within the entire lexicon, but only that among items which become active because they are displayed simultaneously. To test this, we created a small, artificial lexicon with specific lexical similarity characteristics. Subjects learned novel names for 16 novel geometric objects. Objects were presented with high, medium or low frequency during training. Each lexical item had two potential competitors. The crucial comparison was between high-frequency items which had either high- or low-frequency competitors. In spoken word recognition, performance is correlated with the number of frequency-weighted neighbors (phonologically similar words) a word has, suggesting that neighbors compete for recognition as a function of frequency and similarity (e.g., Luce & Pisoni, 1998). We found that in the visual world paradigm, fixation probabilities for items with high-frequency neighbors were delayed compared to those for items with low-frequency neighbors, even when the items were presented with unrelated items. This indicates that fixation probabilities reflect the internal structure of the lexicon, and not just the characteristics of displayed items.

Introduction

Understanding the structure and role of the lexicon in spoken word recognition has implications at higher and lower levels of processing. Recent theories have placed much syntactic, semantic and pragmatic knowledge in the lexicon (e.g., MacDonald, Pearlmutter & Seidenberg, 1994; Tanenhaus & Trueswell, 1995). Thus, representations activated in the resolution of word recognition may have cascading effects which come into play at higher levels. At the same time, lexical knowledge also has effects at lower levels, such as aspects of speech perception which have often been considered pre-lexical (e.g., Andruski, Blumstein, & Burton, 1994; Marslen-Wilson & Warren, 1994). Understanding the structure of the lexicon and lexical activation patterns in spoken word recognition will clearly provide a vital step towards understanding language processing.

A few key parameters have been identified which account for substantial amounts of variability in spoken word recognition. Luce and colleagues (e.g., Luce and Pisoni, 1998) have shown that *log word frequency* alone can account for 4 to 6% of the variance observed in word identification under noise, whereas 16 to 22% of the variance can be accounted for by the *frequency-weighted neighborhood probability rule* (FWNPR), which is the basis of their *Neighborhood Activation Model* (NAM). The FWNPR estimates the amount of expected competition between a word and its "neighbors" (similar words, often defined as words differing by no more than one phoneme), weighted by their frequencies.

These sorts of results inform us about what Marslen-Wilson (1993) has termed the *macrostructure* of spoken word recognition. However, they provide little information about the *microstructure* of the on-line lexical processing -- e.g., what determines the nature of the competitor set over time and the time course of competition effects. Instead, they provide coarse, indirect information. In a typical experiment, recognition and accuracy are measured, but these are usually all-or-nothing data measures of post-recognition decisions, which do not tell us *directly* about on-line processing. Instead, mechanisms of on-line processing must be inferred *indirectly* by seeing how well different parameters (e.g., frequency) correlate with performance.

The interactive-activation connectionist model, TRACE (McClelland and Elman, 1986), is an example of a class of models which provide a different method of testing predictions (an implementation of the Luce and Pisoni, 1998, NAM would be another example). Given a simulated input, TRACE provides a continuous prediction over time of which words in the lexicon should be active and competing for recognition. In the top panel of Figure 1, we present TRACE activations for a target input, *beaker*, an onset competitor (called a "cohort" item because the Cohort model predicts that mainly items which share onsets compete), *beetle*, a rhyme, *speaker*, and an unrelated item (the activations are scaled; see below). But with these fine-grained predictions in hand, how can we test them? Conventional psycholinguistic tasks cannot provide continuous, on-line measures of activation using continuous speech; tasks which are used to try to make time course measurements, such as gating, require interrupting the speech stream and thus using an unnatural stimulus.

Tanenhaus and his colleagues have developed an eye tracking method for studying spoken language comprehension which provides a sensitive, continuous measure of lexical activation (e.g., Tanenhaus et al., 1995).

In this "visual world paradigm", a subject sees a display containing several objects (either real objects or pictures on a computer display). When subjects are asked to perform an action with one of the objects (e.g., "pick up the beaker" or "click on the beetle"), their eye movements are closely time-locked to the speech stream. For example, subjects might be shown a display containing objects *beaker*, *beetle*, *speaker* and *carriage*. If they are asked to "pick up the beaker", the probabilities of fixating each item over time can be compared directly to TRACE predictions.

In the lower panel of Figure 1 are data from Allopenna, Magnuson and Tanenhaus (1998), who presented many such displays to several subjects (the data shown are averaged over several items and several subjects). As can be seen by comparing the upper and lower panels of Figure 1, the data are very similar to the TRACE predictions. Note that the fixation probabilities do not sum to 1 because subjects begin each trial fixating a central fixation cross, and that the TRACE activations have been transformed to simulate the experimental situation of having limited response possibilities (see Allopenna et al. for details).

Note also that while most of the change in probabilities occur after target offset, the fixation probabilities are more closely time-locked to the spoken input than they appear. In very simple tasks, participants require approximately 150 msec to plan and launch a saccade (e.g., Matin, Shao, & Boff, 1993). Allowing for this planning time, it is clear that the earliest eye movements are being planned approximately 100 msec after target onset.

Some other notable qualities of the paradigm are that it does not require subjects to make explicit decisions about stimuli. Instead, eye movements are monitored as subjects respond naturally to continuous spoken instructions. Given a properly constrained task (one in which visually guided movements are required, which allows a functional interpretation of eye movements and avoids the problems identified by Viviani, 1990; see Allopenna et al., 1998, for further discussion), eye movements provide an incidental measure of moment-to-moment attention.

The results reported by Allopenna et al. demonstrate the sensitivity of the visual world paradigm. While cohort (onset overlap) effects were well-established (e.g., Marslen-Wilson & Zwitserlood, 1989), rhyme effects had proven more elusive. For example, weak rhyme effects had been reported in cross-modal and auditory-auditory priming (Connine, Blasko & Titone, 1993; Andruski et al., 1994) only when the rhymes differed by only one or two phonetic features. Allopenna et al.'s rhymes all differed by more than two features. Thus, in addition to providing information about the time course of activation, the visual world paradigm also proved to be more sensitive than other spoken word recognition paradigms.

However, objects whose names were predicted to compete were displayed at the same time. While this allowed the most direct comparison with, e.g., TRACE predictions, the results are ambiguous in one respect: they may have been due to the use of displays with a restricted set of items. That is, competition may have been limited to the set of displayed items, and may not have reflected the influence of the rest of the lexicon. This is important because the

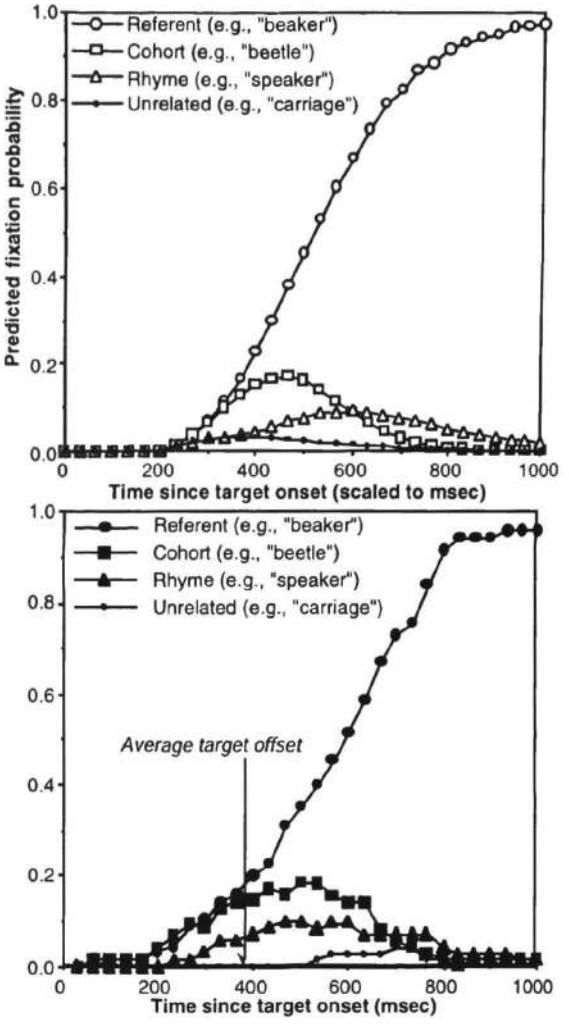

Figure 1: TRACE activations converted to predicted fixation probabilities (top panel) and observed probabilities of fixating a target, a cohort, a rhyme, and an unrelated object from Allopenna et al. (1998).

strength of other paradigms is their ability to inform hypotheses about which lexical items are activated by a given input -- although this is through *indirect* evidence of competition, as reflected in reduced performance. The visual world paradigm provides a relatively *direct* measure of the relative activations of displayed lexical items, but does it indicate the activation of items which are not present?

One way to determine whether competition is limited to the displayed portion of the lexicon is to measure responses to items with different frequency-weighted neighborhood densities (FWNDs). Recall that Luce and colleagues have shown that a word's FWND accounts for much of the variance in spoken word recognition, and consider an example of two words, A and B. If both have 2 neighbors, and all their neighbors have equal occurrence frequencies, recognition times for A and B should be equivalent. If we increase A's FWND by giving it more neighbors, it should take longer to recognize A (because now, given A, more words compete for recognition). If instead we increase the frequency of A's 2 neighbors, it should still take longer to recognize A (since words of higher frequency compete more

strongly, and A's FWND has increased). How would such a neighborhood density effect manifest itself under the visual world paradigm? We could present word A among three unrelated items, and present word B among three unrelated items. If A's FWND is higher, the probability of fixating A over time should increase more slowly than for B.

Here, we report an experiment of just this type, which replicates and extends previous work using an artificial lexicon (Magnuson, Dahan, Allopenna, Tanenhaus and Aslin, 1998). The advantage of using an artificial lexicon is that we can carefully control the statistics of the lexicon. A set of stimuli drawn from a natural language will be more variable. The artificial lexicon lets us test our hypothesis with a minimum of potential confounds. Before turning to the current experiment, we will briefly review the artificial lexicon study on which the current study is based.

Artificial Lexicons and the Visual World Paradigm

In the previous artificial lexicon study (Magnuson et al., 1998), we trained subjects to recognize a lexicon of 16 novel words. Each lexical item (e.g., /pibo/) had two potential competitors: a cohort (e.g., /pibu/) and a rhyme (e.g., /dibo/). Each item in the lexicon was randomly associated with a novel geometrical object. Subjects learned the lexicon by learning the names for each object.

Figure 2 shows examples of the sorts of displays subjects saw on a computer screen. Initially, subjects saw pairs of objects, and heard instructions to click on one with the computer mouse (e.g., "click on the pibo"). At first, subjects had to guess. But after they clicked on one object, they received feedback: one object would disappear, and they knew that the remaining one was being named, and then they would hear the name again. Different levels of word frequency were approximated by presenting the items with "high" or "low" frequency (with a ratio of 7:1/high:low during training). Item frequency was crossed with competitor frequency: four items were high frequency (HF) and had HF neighbors (HF/HF); four were low frequency (LF) with low frequency neighbors (LF/LF); four were HF/LF, and four were LF/HF.

Subjects quickly reached ceiling on the 2AFC (alternative forced choice) task, and training continued with a 4AFC task (see Figure 2). After 80 minutes of training on each of two days, we monitored eye movements as subjects performed the basic visual world paradigm task without feedback: given

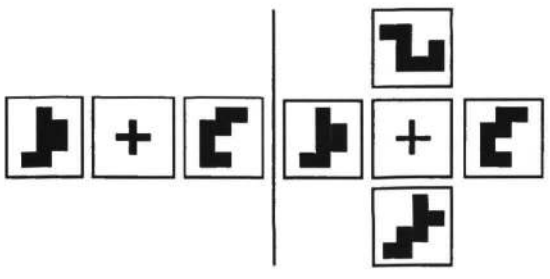

Figure 2: Examples of stimulus displays. The left panel shows a possible display in 2AFC training; the right panel shows a possible 4AFC display.

a display containing four objects, subjects were instructed to click on one of the objects. On most trials, the items were all unrelated. On critical trials, a cohort or rhyme competitor was present. The results after two days closely resembled Allopenna et al.'s (1998) results using real words: there was cohort and rhyme activation, and the fixation probabilities for each object varied as a function of the similarity of the stimulus with the object's name. There were also interactions between item frequency and competitor frequency. For example, the cohort effect was stronger -- with a substantial initial advantage for the cohort -- for low-frequency items with high-frequency competitors than for high-frequency items with high-frequency competitors. This indicates that, as with real words, neighbors competed as a function of their similarity and frequency.

The experiment included a FWND manipulation (although all items had the same number of neighbors, FWND varied because competitor frequency varied), and a condition where items were presented along with three unrelated distractors. However, this condition did not provide a complete test of the question at hand: namely, whether changes in fixation probability over time reflect activation of present and absent competitors. This is because there were only two levels of frequency, and target items were presented among unrelated distractors which were matched in frequency with the target's competitors (e.g., for a high-frequency target with low-frequency competitors, low-frequency, unrelated distractors were used). Thus, we cannot be certain that differences in fixation probabilities were due to the frequencies of the absent competitors, or to the frequencies of the simultaneously presented distractors.

In order to have a clean test of the hypothesis, we need a third level of frequency. Then, HF/HF and HF/LF items can both be presented among unrelated distractors of the same frequency, and differences we observe should be due to the frequencies of (absent) competitors, not the characteristics of the unrelated distractors. This was the design we used for the current experiment.

Absent Competitors and the Visual World Paradigm

The design of the current study was similar to that used by Magnuson et al. (1998). We trained subjects to recognize a lexicon of 16 novel words. Each lexical item (e.g., /pibo/) had two potential competitors: a cohort (e.g., /pibu/) and a rhyme (e.g., /dibo/), and was randomly associated with a novel geometrical object. Target and competitor frequency were varied, but with three levels rather than two. The third level (medium) provided distractors of uniform frequency to serve as distractors for the other items.

Method

Participants Seven students at the University of Rochester were paid for their participation. All were native speakers of English with normal or corrected-to-normal vision and normal hearing.

Materials The visual stimuli were simple patterns, formed by filling eight randomly-chosen, contiguous cells of a four-by-four grid (see Figure 2). 10,000 such randomly-generated patterns were randomly ordered, and sixteen were selected

from the beginning of the set (with two items replaced due to visual similarity with other items).

The novel words consisted of sixteen bisyllabic nonsense words. The sixteen words comprised four four-word sets, such as /pibo/, /pibu/, /dibo/, and /dibu/. Note that for each word, there is an onset ("cohort") competitor which differs only in the final vowel, a rhyme, and a relatively dissimilar item (differing by two phonemes, which would not qualify it as a neighbor using the most standard definition of a word differing by a single phoneme). A small set of phonemes was selected in order to achieve consistent similarity within and between sets. The consonants /p/, /b/, /t/, and /d/ were chosen because they are among the most phonetically similar stop consonants. In each set, rhymes differed by two phonetic features (place and voicing) in the first phoneme. Transitional probabilities were controlled such that all phonemes and combinations of phonemes were equally predictive at each position or combination of positions.

The auditory stimuli were produced by a male, native speaker of English in a sentence context ("click on the ____"). The average duration of the target words was 496 msec. The stimuli were recorded to tape, and then digitized using the standard analog/digital devices on an Apple Macintosh 8500 at 16 bit, 44.1 kHz. The stimuli were converted to 8 bit, 11.127 kHz (SoundEdit format) in order to be used with the experimental control software, PsyScope 1.2 (Cohen, MacWhinney, Flatt & Provost, 1993).

Procedure Participants came to the lab for two 2-hour sessions on consecutive days. Each day consisted of seven training blocks with feedback and a testing block without feedback. We tracked eye movements during the test.

Participants were seated at a comfortable distance from the experimental control computer (an Apple Macintosh 7200 PowerPC). The structure of the training blocks was as follows. First, a fixation cross would appear on the screen. The participant had to click on the cross to begin the trial. After 500 msec, either two shapes (in the first four training blocks) or four shapes (in the rest of the training blocks and the tests) would appear. If only two shapes were presented, they appeared at about 1.5 degrees of visual angle to the left and right of the fixation cross. When four shapes were presented, two would also appear about 1.5 degrees above and below the fixation cross (see Figure 2).

Participants heard the instruction, "Look at the cross" through headphones 750 msec after the objects appeared. Then, they fixated the cross and clicked on it. Participants were instructed at the beginning of the session that they should fixate the cross until they heard the next instruction. 500 msec after clicking on the cross, an instruction to click on one of the items (with the computer's mouse) was presented (e.g., "Click on the pibu").

When participants responded by clicking on one of the items, or at the end of 15 seconds, all of the items disappeared except for the shape that was actually named. The correct shape's name was repeated 500 msec later. The object disappeared 500 msec later, and the subject would click on the cross to begin the next trial. The testing block was identical to the four-item training, except that no feedback was given.

Shapes were randomly mapped to names, with a different random mapping for each subject. Half the items were medium frequency. Six items were high frequency, and two were low frequency. All of the medium frequency items had medium frequency competitors. The high- and low-frequency items were assigned such that four of the high frequency items had high frequency competitors, and two of the high frequency items had low frequency competitors (and the competitors for the two low frequency items were those two high frequency items).

Each training block consisted of 68 trials. High frequency items appeared 7 times per block, low-frequency items appeared once per block, and medium frequency items appeared 3 times per training block. Across all training blocks, all items appeared as visual distractors approximately equally often. Within training, distractors were randomly assigned to each trial.

The tests consisted of 96 trials. Each item appeared in six trials: one with its onset competitor and two unrelated items, one with its rhyme competitor and two unrelated items, and four with three unrelated items. For the crucial comparisons (HF/HF and HF/LF), medium frequency items were used as unrelated distractors.

We tracked eye-movements with an Applied Scientific Laboratories (E4000) eye tracker. Two cameras mounted on a lightweight helmet provide the input to the tracker. The eye camera provides an infrared image of the eye. The center of the pupil and the first Purkinje corneal reflection are tracked to determine the position of the eye relative to the head. Accuracy is better than 1 degree of arc, with virtually unrestricted head and body movements. A scene camera is aligned with the participant's line of sight. A calibration procedure allows software running on a PC to superimpose crosshairs showing the point of gaze on a HI-8 video tape record of the scene camera. The scene camera samples at a rate of 30 frames per second, and each frame is stamped with a time code. The auditory stimuli were presented binaurally through headphones using the standard digital-to-analog devices provided with the experimental control computer. Audio connections between the computer and HI-8 VCR provided an audio record of each trial. Each trial was analyzed frame-by-frame from stimulus onset to the subject's response (clicking on the appropriate object) by coding fixations from each saccade onset.

Results

Subjects were able to achieve high levels of accuracy relatively quickly; accuracy for high, medium and low frequency items was .83, .89, and .77, respectively, on the 4AFC test without feedback on day 1, and .96, .97, and .94 on day 2. Figure 3 demonstrates that we replicated the basic cohort and rhyme effects reported by Allopenna et al. (1998) and Magnuson et al. (1998). The results are averaged over all conditions in which cohort or rhyme competition was possible.

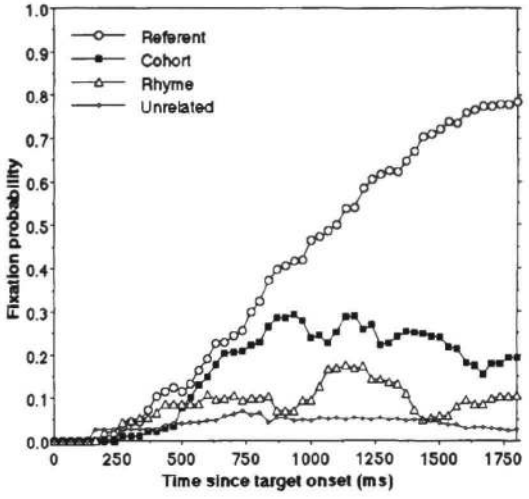

Figure 3: Combined cohort and rhyme effects.

Figure 4 shows how the cohort effect is modulated by target and competitor frequency. In the top panel, the target is low-frequency, and the cohort is high, and there was an initial advantage for the cohort. In the middle, both target and cohort were high frequency, and the result resembles the cohort effect shown in Figure 1, with no advantage for either item initially. On the bottom, the target was high- and the cohort was low-frequency. Although we do not see the expected initial advantage for the target, the cohort clearly is less active than in the other two panels. Thus, the results replicate Magnuson et al. (1998) and show that the degree to which (simultaneously present) items compete depends both on their similarity to the input, and their frequency -- as predicted by models such as NAM and TRACE.

Figure 5 shows the crucial comparison between HF/HF and HF/LF items presented among three unrelated, medium frequency distractors. The frequencies of the targets are equal, and the unrelated distractors are matched in the two cases. The only difference between the items is the frequency of their *absent* competitors. As predicted, the probability of fixating the HF/HF target increases much more slowly than the probability of fixating the HF/LF target. In a 2-way ANOVA (frame x competitor frequency) on the two referent probabilities from frame 10 (333 ms, when the probabilities first diverge) to frame 45 (1500 ms), both main effects were reliable (frame: $F(35, 210) = 30.97$, $p < .001$; competitor frequency: $F(1,6) = 9.00$, $p < .05$). A paired, one-tailed t-test on average fixation probabilities over the same window was also significant ($t(6) = 1.98$, $p < .05$, mean difference = .103).

Discussion

The current results show that fixation probabilities in the visual world paradigm reflect lexical competition which includes items which are not visually present. As in tasks such as lexical decision, recognition time depends on the internal structure of the lexicon. Since the only difference between the HF/HF and HF/LF targets is the frequency of their competitors, we see that lexical items for which no visual referents are available still compete for recognition.

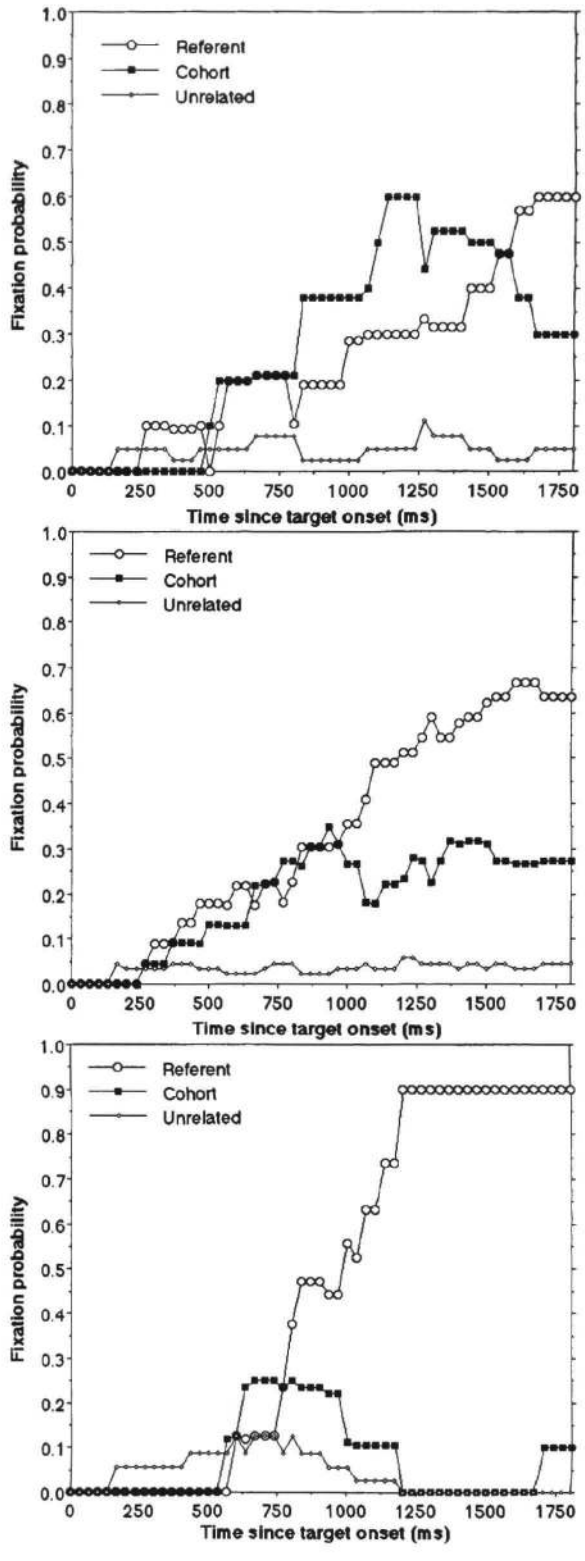

Figure 4: Cohort effects as a function of target and competitor frequency. Top: low-frequency target, high-frequency cohort. Middle: high-frequency target and cohort. Bottom: high-frequency target, low-frequency cohort.

The data in Figure 5 allow us to disregard an alternative interpretation of the results shown in Figure 4, where target

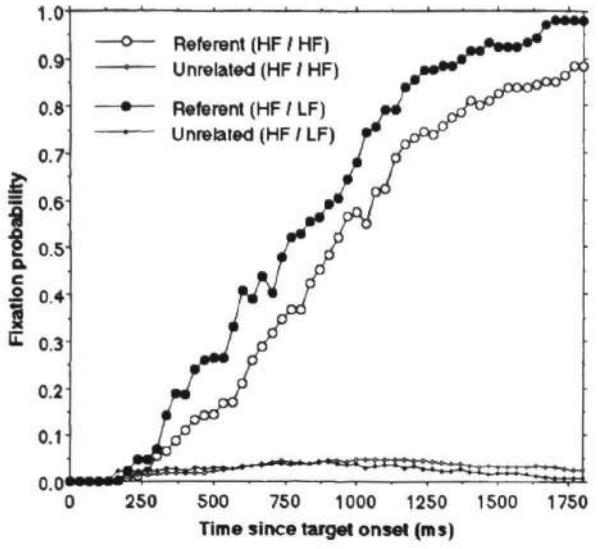

Figure 5: Target fixation probabilities as a function of competitor frequency. Both targets were high frequency, and were presented with 3 unrelated distractors of medium frequency. The only difference between the targets was the frequency of their competitors, *which were not displayed.*

probabilities rise most quickly when the target is high frequency and the competitor is low frequency. We might infer that this indicates competition between the items. However, fixations are necessarily serial. Given what appears to be competition among a set of simultaneously displayed items, one cannot infer competition at the lexical level, rather than simple co-activation (when data is sparse). The site of the competition could be the motor programming to move the eye. Once an item is fixated, the subject cannot simultaneously indicate the activation of other objects. An increased fixation probability for one item may be accompanied by a reduced probability for another, which could lead to the appearance of lexical competition. This problem of interpretation is diminished with sufficiently many data points. Despite having data from relatively few subjects, the current results can be interpreted as indicating lexical competition rather than competition at fixation generation, since the differences in target fixation probabilities shown in Figure 5 are not accompanied by commensurate differences in unrelated fixation probabilities. Therefore, the differences indicate that more time was needed for the activation of the target to become sufficiently large to generate initial eye movements when the target had high-frequency competitors.

In summary, we have replicated the results of Magnuson et al. (1998), showing an interaction of target and competitor frequency. This work also shows that effects in the visual world paradigm are not driven simply by competition among visible referents; changes in fixation probabilities are also driven by competition for recognition with competitors which are not visually available.

Acknowledgments

Supported by NIH HD27206 to MKT, NSF SBR-9729095 to MKT and RNA, and an NSF Graduate Research Fellowship to JSM.

References

Allopenna, P. D., Magnuson, J. S., & 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.

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition*, 52, 163-187.

Cohen J. D., MacWhinney B., Flatt M. & Provost J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Res. Methods, Instruments & Computers*, 25(2), 257-271.

Connine, C. M., Blasko, D. G., & Titone, D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition? *J. Memory & Language*, 32, 193-210.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The Neighborhood Activation Model. *Ear and Hearing*, 19, 1-36.

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

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. *Proc. 20th Annual Conference of the Cognitive Science Society*, 651-656.

Marslen-Wilson, W. (1993). Issues of process and representation in lexical access. In G. T. M. Altmann & R. Shillcock (Eds.), *Cognitive Models of Speech Processing: The Second Sperlonga Meeting*. Erlbaum.

Marslen-Wilson, W., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psych. Rev.*, 101, 653-675.

Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, 15, 576-585.

Matin, E., Shao, K. C., and Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception & Psychophysics*, 53, 372-380.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psych.*, 18, 1-86.

Tanenhaus, M. K., and Trueswell, J. C. (1995). Sentence comprehension. In J. L. Miller & P. D. Eimas (Eds.), *Handbook of Perception and Cognition, Volume 11: Speech, Language and Communication*. San Diego: Academic Press.

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

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.