Journal of Experimental Psychology: General
2003, Vol. 132, No. 2, 202–227

0096-3445/03/$12.00 DOI: 10.1037/0096-3445.132.2.202

The Time Course of Spoken Word Learning and Recognition: Studies With Artificial Lexicons

James S. Magnuson
Columbia University

Michael K. Tanenhaus and Richard N. Aslin
University of Rochester

Delphine Dahan
Max Planck Institute for Psycholinguistics

The time course of spoken word recognition depends largely on the frequencies of a word and its competitors, or *neighbors* (similar-sounding words). However, variability in natural lexicons makes systematic analysis of frequency and neighbor similarity difficult. Artificial lexicons were used to achieve precise control over word frequency and phonological similarity. Eye tracking provided time course measures of lexical activation and competition (during spoken instructions to perform visually guided tasks) both during and after word learning, as a function of word frequency, neighbor type, and neighbor frequency. Apparent shifts from holistic to incremental competitor effects were observed in adults and neural network simulations, suggesting such shifts reflect general properties of learning rather than changes in the nature of lexical representations.

Current models of spoken word recognition share a set of core assumptions that correspond to what Marslen-Wilson (1993) called the *macrostructure* of spoken word recognition: As speech is heard, multiple lexical candidates are activated and compete for recognition with strengths proportional to their similarity with the input and their prior probabilities (frequencies of occurrence). Perhaps the best evidence for multiple activation and "proportional" competition comes from studies by Luce and colleagues (e.g., P. A. Luce & Pisoni, 1998; P. A. Luce, Pisoni, & Goldinger, 1990), who found that recognition in a variety of tasks depends on not just the frequency of a target word but also its *frequency weighted neighborhood*: the summed log frequencies of words expected to compete with the target based on a phonetic similarity metric. The fact that recognition difficulty is proportional to neighborhood density suggests that as a word is heard, each neighbor is activated and competes for recognition. Although these results provide general constraints on models of spoken word recognition, as Marslen-Wilson (1993) argued, a deeper understanding requires empirical explorations of *microstructure*, addressing questions such as precisely which items are activated in response to a given input, the time course of competition resolution, and detailed specification of the mechanisms underlying word recognition.

The answer to many of these questions can only be obtained by examining whether lexical access in continuous speech is predicted by the details of lexical neighborhoods (i.e., the number and nature of similar-sounding words expected to compete with one another). However, the temporal characteristics of speech introduce some difficult methodological challenges in defining and controlling the properties of lexical neighborhoods. Some of these challenges can be highlighted by comparing spoken word recognition with visual word recognition, in which lexical neighborhoods have also featured prominently in experimental studies and computational models (e.g., Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989).

In visual word recognition, word boundaries are clearly marked, and one can assume, as a first approximation, that all of the sensory input for a word is simultaneously available. On the assumption that lexical processing abstracts away from details of surface form, such as type font, electronic corpora can be used to provide detailed analyses of lexical neighborhoods. This allows precise control over the details of lexical neighborhoods, facilitating computational explorations and experimental studies.

In contrast, spoken words are composed of a series of transient acoustic events, with no known invariant cues to boundaries between spoken words (Aslin, Woodward, LaMendola, & Bever, 1996; Cole & Jakimik, 1980; Lehiste, 1970). This means similarity between lexical representations and speech input must be mapped incrementally as the input unfolds over time, with word boundaries

James S. Magnuson, Department of Psychology, Columbia University; Michael K. Tanenhaus and Richard N. Aslin, Department of Brain and Cognitive Sciences, University of Rochester; Delphine Dahan, Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands.

This study was supported by National Science Foundation (NSF) Grant SBR-9729095 and National Institute of Child Health and Human Development Grant DC-005071 to Michael K. Tanenhaus and Richard N. Aslin, an NSF Graduate Research Fellowship, a Grant-in-Aid of Research from the National Academy of Sciences, through Sigma Xi, and National Institute on Deafness and Other Communication Disorders Grant DC-005765 to James S. Magnuson. We thank Craig Chambers, Bob McMurray, and George Pappas for helpful discussions and Dana Subik and Gina Cardillo for help in running participants and manuscript preparation.

Correspondence concerning this article should be addressed to James S. Magnuson, Department of Psychology, Columbia University, 1190 Amsterdam Avenue, MC 5501, New York, New York 10027. E-mail: magnuson@psych.columbia.edu

inferred on the basis of possible matches between lexical items and the input. Thus, response measures must be sensitive to the time course of processing as a spoken word unfolds and must allow the experimenter to monitor lexical access in continuous speech.

Models of spoken word recognition are also dependent on the quality of corpus-based measures, and considerable progress has been made using this approach. For example, word frequency estimates from written corpora account for about 5% of the variance in spoken word recognition in tasks such as lexical decision and naming, and P. A. Luce & Pisoni's (1998) phonetic similarity metric accounts for about 20%. Nonetheless, there are several impediments to the corpus-based approach with speech. Most currently available corpora are either based on written sources (e.g., Francis & Kucera, 1982) or consist of automatic transcriptions of spoken corpora to orthographic or canonical phonemic forms (e.g., transcriptions of the CALLHOME corpus; see http://www.ldc.upenn.edu).[1] The most serious limitation of these analyses is that the forms listed in a corpus will differ substantially from the context-dependent surface realization of spoken words in fluent speech. Because the acoustic sound pattern of a phone varies depending on (among other factors) phonetic context, sentential context, talker characteristics, and acoustic conditions, the production of a word in fluent speech may bear little similarity to its canonical "citation" realization. For example, a word produced in one sentence context by a particular talker may differ substantially from the same talker's realization of the same word in a different context (Fougeron & Keating, 1997) or may depend on whether a content word represents old or new discourse information (Fowler & Housum, 1987; Nooteboom & Kruyt, 1987; Terken & Hirschberg, 1994). Subphonemic differences in phonetic realization affect lexical processing (Andruski, Blumstein, & Burton, 1994; Dahan, Magnuson, Tanenhaus, & Hogan, 2001; Marslen-Wilson & Warren, 1994; McMurray, Tanenhaus, & Aslin, 2002; McQueen, Norris, & Cutler, 1999), and various other surface characteristics influence the processing of, and memory for, spoken words (Goldinger, 1996, 1998).

The goal of the present research is to overcome these stimulus control problems by examining factors that determine the time course of lexical activation and competition using an artificial lexicon with precisely controlled distributional and acoustic/phonetic properties. Artificial linguistic materials have been successfully used to provide precise control over distributional information in several domains of language acquisition and processing (e.g., Braine, 1963; Gomez & Gerken, 2000; Morgan, Meier, & Newport, 1987; Saffran, Newport, & Aslin, 1996). In spoken word recognition research, studies with artificial lexicons could complement traditional studies by allowing researchers to create stimuli to test precise hypotheses while greatly reducing the set of uncontrolled variables. Once hypotheses have been tested under the tightly controlled conditions afforded by an artificial lexicon, they can be extended to experiments using real words. An experimenter armed with predictions from an artificial lexicon study would be better able to tease apart effects of experimental manipulations from confounding influences.

In the present studies, participants learned new "words" by associating them with novel visual shapes. We then used the words and shapes in an eye-tracking paradigm that has been shown to provide a sensitive measure of the time course of lexical activation in spoken word recognition (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; see Cooper, 1974, for a related precedent). In this paradigm, participants follow spoken instructions to perform visually guided actions (e.g., picking up real objects or using a computer mouse to click on pictures in a display). In the present experiments, on each trial, participants saw two or four geometric forms on a computer display. Their task was to learn the novel names for these novel objects. They did this by responding to a spoken instruction to click on one of the objects (e.g., "click on the /pibo/"). Feedback informed them as to whether they had selected the correct item, and participants quickly learned the objects' names (more details are provided below). Following training, we tracked the participants' eye movements as they performed this task without feedback and used the eye movements to make inferences about the time course of spoken word recognition.

Requiring participants to perform visually guided movements establishes a functional link between eye movements and the speech stimulus: A participant can perform the motor task more efficiently with visual guidance, and the information specifying the task to perform must be extracted from the spoken instruction. Indeed, eye movements are closely time-locked to speech, and the paradigm has been used to obtain fine-grained time course measures of lexical activation and competition (Allopenna, Magnuson, & Tanenhaus, 1998; Dahan, Magnuson, & Tanenhaus, 2001; Dahan, Magnuson, Tanenhaus, & Hogan, 2001).

The present studies were designed to determine whether an artificial lexicon approach is feasible and to examine at the same time more complex combinations of the factors that determine the time course of lexical competition. We examined four questions. First, would word recognition in an artificial lexicon exhibit the signature processing effects that characterize spoken word recognition with real words? Specifically, would listeners process the input incrementally, activating multiple lexical candidates that then compete for recognition? To address this question, Experiments 1 and 2 examined the processing of high-frequency (HF) and low-frequency (LF) words with HF and LF (cohort) competitors that either differed in only their final phoneme (e.g., /pibo/ and /pibu/) or (rhyme) competitors that differed in only their initial phoneme (e.g., /pibo/ and /dibo/). We examined frequency because ease of recognition depends in part on frequency of occurrence (Connine, Titone, & Wang, 1993; Dahan, Magnuson, & Tanenhaus, 2001; Howes, 1957; Savin, 1963). We chose cohort and rhyme competitors because there is comparable eye-tracking data using real words with which we can compare our results (Allopenna et al., 1998; Dahan, Magnuson, & Tanenhaus, 2001). All current models predict clear effects of cohort competitors, providing an essential criterion for evaluating the artificial lexicon data. In contrast, models differ in their predictions about rhyme competition, allowing us to address a more theoretically controversial issue about the details of lexical neighborhoods.

Although numerous studies using a variety of conventional psycholinguistic tasks have found evidence for cohort competition

[1] Cairns, Shillcock, Chater, and Levy (1995) attempted to compensate for the lack of surface detail in corpora by applying a set of phonological rules to a corpus transcribed with canonical phonemic forms. The resulting corpus would provide a substantial leap toward realistic input compared with the canonical transcriptions but would still vastly underestimate the amount of variability that occurs in natural speech.

(e.g., Marslen-Wilson, 1993; Marslen-Wilson & Zwitserlood, 1989), evidence for rhyme activation has been weak and has typically been found only when items differ by one or two phonetic features (Andruski et al., 1994; Connine, Blasko, & Titone, 1994; Marslen-Wilson, 1993). Allopenna et al. (1998) tested the possibility that rhyme competition escaped detection in conventional tasks because it is relatively weak, as predicted by the TRACE model (McClelland & Elman, 1986). Allopenna et al. found evidence for rhyme competition that mapped onto the predictions from TRACE: Participants fixated rhymes reliably more than unrelated items but also substantially less than cohorts (and the time course of fixations to cohorts and rhymes mapped onto emerging phonetic similarity as target words were heard).

TRACE differs from Shortlist (Norris, 1994) and the Cohort model (Marslen-Wilson, 1987, 1993) in the way the bottom-up signal is mapped onto lexical representations. In Cohort and Shortlist, bottom-up mismatch detection prevents the bottom-up input from activating items with high overall similarity when they have substantial onset dissimilarity (e.g., rhymes).[2] In TRACE, there is no bottom-up mismatch detection, and words can be activated by overlap with any part of the input. Rhymes, however, are still at a disadvantage because lexical selection is achieved in part through lateral inhibition between lexical nodes. A disadvantage for rhymes compared with items overlapping at onset emerges because items with onset similarity to an input are activated earlier than rhymes, with the effect that they inhibit rhymes. The two sorts of models differ in a subtle fashion (depending on parameter settings): TRACE predicts late, weak rhyme activation, even given substantial onset mismatch, whereas models like Cohort and Shortlist predict late, weak rhyme activation only when rhymes differ minimally.

The present experiments allow us to compare cohort and rhyme competition. Importantly, they also allow us to evaluate how the dynamics of lexical access change with degree of learning. It has been argued that children's lexical representations may be substantially different from adults, with children having more holistic lexical representations (e.g., syllable-based rather than segment-based) than adults (e.g., Charles-Luce & Luce, 1990; Ferguson & Farwell, 1975). For example, positional overlap might be less important than overall similarity early on in learning because the acoustic/phonetic specificity of the lexical items may be coarsely coded (i.e., if neighborhoods are sparse, relatively less phonetic detail might be required to adequately match a lexical item, or children's segmental representations may be less refined). By studying competition effects as adults learn new sets of words, we can ask whether adults show any evidence of "holistic" representations. Specifically, we can examine whether cohorts compete more strongly than rhymes even early in training. Incremental competition effects early in training would support the notion that adults and children have fundamentally different lexical representations. However, if we find that adults show patterns consistent with a holistic-to-incremental shift over learning, the shift may be evidence of a general phenomenon associated with early learning rather than a fundamental developmental change in the nature of lexical representations.

The second question we examined was whether word recognition in the eye-tracking paradigm shows effects of the entire artificial lexicon. When possible words are unconstrained, participants in conventional experiments may expect to hear any lexical item, and any lexical representation could be activated, given adequate similarity to the input. In the eye-tracking paradigm, a limited response set is used; participants choose from a small set of possible referents (pictures), raising the possibility that contextual expectations might constrain which lexical items are activated during speech processing. Experiment 2 tested whether word recognition in the artificial lexicon paradigm is based on open-ended lexical activation or if activation is limited to the displayed alternatives.

The third question was whether the present experiments could be modeled using a learning architecture with a well-defined linking hypothesis to on-line changes in fixation. In previous work (Allopenna et al., 1998; Dahan, Magnuson, & Tanenhaus, 2001; Dahan, Magnuson, Tanenhaus, & Hogan, 2001; for a review, see Tanenhaus, Magnuson, Dahan, & Chambers, 2000), we have developed an explicit linking hypothesis between eye movement data in spoken language understanding tasks and predictions from TRACE (McClelland & Elman, 1986). One limitation of working with TRACE is that it is an interactive activation model with fixed connection weights, which makes it unsuitable for examining learning. We examined whether a simple recurrent network model, or SRN (a connectionist model that is ideally suited to learning temporal dependencies; e.g., Elman, 1990), could account for the effects of phonetic similarity and frequency on lexical access and its development. If such a model captures the dynamics of lexical activation and competition during and after learning, this would suggest SRNs might prove useful for developing models that could account for both learning and processing in the native-language lexicon.

Finally, we asked whether lexical processing in an artificial lexicon would show competition effects from words in a participant's native lexicon. Intrusion may be likely, given evidence for cross-linguistic phonological competition in bilinguals (Spivey & Marian, 1999). However, participants quickly adapt to experiment-specific subsyllabic patterns (Chambers, Onishi, & Fisher, 2003; Dell, Reed, Adams, & Meyer, 2000; Onishi, Chambers, & Fisher, 2002), and we might expect little intrusion from English during the experiment. Large intrusion effects would undermine the goal of creating a lexicon with well-understood properties, thus limiting the usefulness of the paradigm. Experiment 3 addressed this issue by examining the processing of HF and LF novel words that were constructed to fall into either dense or sparse English neighborhoods.

Experiment 1

The artificial lexicon used in this experiment consisted of 16 novel bisyllabic names, each of which was associated with a different novel shape. Each name had one onset competitor and one rhyme competitor. Neighborhood was held constant (all words had neighborhoods of three: the word itself, e.g., /pibo/, plus an onset competitor and a rhyme, e.g., /pibu/ and /dibo/). Word frequency was manipulated by presenting 8 of the words with

[2] In the case of either Cohort or Shortlist, degree of mismatch sensitivity is a modifiable parameter. Either model could be made more similar to TRACE by relaxing the weight of bottom-up mismatches or adjusting the featural threshold for mismatch detection. However, an explicit assumption of both models is that bottom-up mismatch detection will provide an optimal parsing of phonemic input to the word recognition system.

relatively high frequency and the other 8 with relatively low frequency. In addition, 4 of the HF items had LF competitors, and 4 had HF competitors. The same was true for the LF items. These manipulations were introduced to allow us to examine how competition effects were modulated by frequency.

The experiment examined three questions. First, would word recognition in the artificial lexicon eventually show the pattern of incremental interpretation that characterizes skilled word recognition? We addressed this question by comparing the results from the artificial lexicon with comparable manipulations of real words from Allopenna et al. (1998). Second, how does frequency modulate cohort and rhyme competition? This was addressed by comparing the effects of frequency on cohort competition, such as those reported by Dahan, Magnuson, and Tanenhaus (2001), with similar conditions in the artificial lexicon. In addition, the artificial lexicon allowed us to evaluate whether frequency would modulate neighbor competition effects, as predicted by the neighborhood activation model (P. A. Luce & Pisoni, 1998). The third question was how incremental processing develops during learning. There are claims in the developmental literature that children's early lexical representations are more holistic than adults', for example, due to their being based on syllables rather than phonemic segments (e.g., Ferguson & Farwell, 1975; Garnica, 1973; Menyuk & Menn, 1979; Shvachkin, 1948/1973). If global similarity is more important than positional overlap early in learning, then rhyme and cohort effects might be equivalent early in learning, with stronger cohort effects emerging only after more exposure.

Method

Participants. Sixteen students at the University of Rochester who were native speakers of English with normal hearing and normal or corrected-to-normal vision were paid $7.50/hr for participation.

Materials. The visual stimuli were simple patterns formed by filling eight randomly chosen, contiguous cells of a 4 × 4 grid (i.e., a "starter cell" was chosen randomly, and then seven more in series, with the constraint that each be above, below, left of, or right of the preceding selection; see Figure 1). Pictures were randomly mapped to words.[3] The artificial lexicon consisted of 16 bisyllabic novel words, organized into four 4-word sets, such as /pibo/, /pibu/, /dibo/, and /dibu/.[4] The words were pronounced such that the syllable boundary was between the first vowel and second consonant. Within each 4-word set, each word had an onset-matching (cohort) neighbor, which differed only in the final vowel; an onset-mismatching (rhyme) neighbor, which differed only in its initial consonant; and a dissimilar item, which differed in the first and last phonemes. The cohorts and rhymes qualify as neighbors under the "short-cut" neighborhood metric of items differing by a one-phoneme addition, substitution, or deletion (e.g., Newman, Sawusch, & Luce, 1997), which works nearly as well as the more complex metrics Luce and colleagues (e.g., P. A. Luce & Pisoni, 1998) have devised. A small set of phonemes was selected 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. The first phonemes of rhyme competitors differed by two phonetic features: place and voicing. Transitional probabilities were controlled such that all phonemes and combinations of phonemes were equally predictive at each position and combination of positions. All of the artificial lexical items would fall into sparse English neighborhoods (with 0–2 neighbors [$M = 0.6$] and frequency-weighted neighborhood densities [summed log frequencies of neighbors[5]] ranging from 0 to 5.9 [$M = 1.9$]), and all would have relatively few cohort competitors (English words overlapping with them in the initial CV [consonant–vowel]; the range was 7–23, and the mean was 12.5) and relatively low frequency-weighted cohort densities (computed in analogy to the neighborhood probability, i.e., the summed log frequencies of words overlapping in the initial CV; the range was 9.5–63.5, with a mean of 32.2). Note that all of the items would fall into the low-cohort density category of English materials used by Magnuson (2001) in a comparison of cohort and neighborhood density. For that study, the low-density cohort items had from 5 to 67 cohorts and cohort densities from 6.4 to 98.1 ($M = 47.3$), whereas the high-cohort density items had from 84 to 325 cohorts and cohort densities from 152.3 to 975.5 ($M = 289.0$). Thus, we would expect little effect of English lexical items on the artificial items, but if there were, we would not expect any

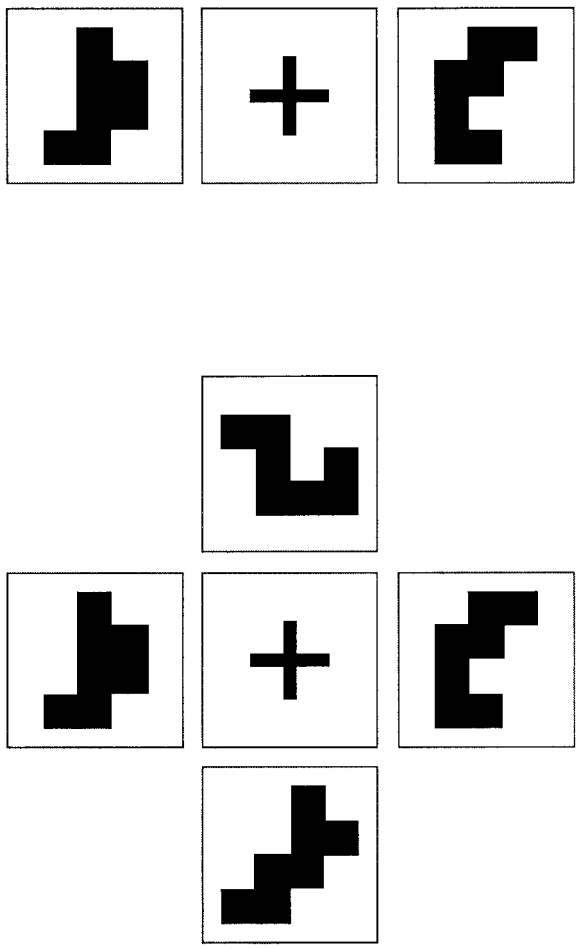

Figure 1. Examples of stimulus displays. An example of a two-alternative forced choice display is shown above; a four-alternative forced choice example is shown below.

[3] Two random mappings were used for the first 8 participants, with 4 assigned to each mapping. A different random mapping was used for each of the 8 participants in the second group. Analyses of variance (ANOVAs) using group as a factor showed no reliable differences, so we have combined the groups.

[4] The four sets were (a) /pibo/, /pibu/, /dibo/, /dibu/; (b) /pota/, /poti/, /dota/, /doti/; (c) /bupa/, /bupi/, /tupa/, /tupi/; and (d) /bado/, /badu/, /tado/, /tadu/.

[5] Note that the frequencies and densities we report are based on natural logs; apparent differences in levels compared with some other studies may be due to their use of base 10 logs.

items to be selectively affected (in Experiment 3 we address the issue of the influence of the native lexicon directly).

The auditory stimuli were produced by a phonetically trained male native speaker of English in a sentence context ("Click on the pibo."). 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) for use with the experimental control software, PsyScope 1.2 (Cohen, MacWhinney, Flatt, & Provost, 1993). The mean duration of the "click on the . . . " portion of the instruction was 456 ms, and the mean duration of the artificial lexical items was 496 ms.

Procedure. Participants were trained and tested in two 2-hr sessions on consecutive days. Each day consisted of seven training blocks with feedback and a test without feedback. Eye movements were tracked during the test.

The structure of the training trials was as follows. First, a central fixation cross appeared on the screen. The participant then clicked on the cross to begin the trial. After 500 ms, either two shapes (in the first three training blocks) or four shapes (in the rest of the training blocks and the tests) appeared (see Figure 1). Participants heard the instruction, "Look at the cross," through headphones 750 ms after the objects appeared. Participants fixated the cross, then clicked on it with the mouse, and continued to fixate the cross until they heard the next instruction (as they had been instructed to do prior to the start of the experiment). After 500 ms of clicking on the cross, the spoken instruction was presented (e.g., "Click on the pibu."). When participants responded, all of the distractor shapes disappeared, leaving only the correct referent (again, as participants knew from preexperiment instruction). The name of the shape was then repeated. The object disappeared 500 ms later, and the participant clicked on the cross to begin the next trial. The test was identical to the four-item training, except that no feedback was given (150 ms after the participant clicked on an item, all four items disappeared).

During training, half of the items were presented with HF and half with LF. Half of the eight HF items had LF neighbors (e.g., /pibo/ and /dibu/ might be HF, and /pibu/ and /dibo/ would be LF), and vice versa. The other items had neighbors of the same frequency. Thus, there were four combinations of word/neighbor frequency: HF/HF, LF/LF, HF/LF, and LF/HF. Each training block consisted of 64 trials. HF items appeared as targets seven times per block, and LF items appeared once per block as targets. Each item appeared in six test 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 (96 total).[6] The distractors for each training trial were randomly selected for the two-alternative forced choice (2AFC) training, as well as for the four-alternative forced choice (4AFC) training, except for the constraint that at least one item had to be from a different four-word set than the target.[7] Position in the display was random on each trial for both types of training and the test.

Eye movements were monitored using an Applied Sciences Laboratories E4000 eye tracker, which provided a record of point-of-gaze superimposed on a video record of the participant's line of sight (see Tanenhaus & Spivey-Knowlton, 1996). The auditory stimuli were presented binaurally through headphones using standard Macintosh Power PC digital-to-analog devices and simultaneously to the HI-8 VCR, providing an audio and video record of each trial. Trained coders (blind to picture–name mapping and trial condition) recorded eye position within one of the cells of the display at each video frame.

Results

A response was scored as correct if the participant clicked on the named object with the mouse. Participants were close to ceiling for HF items in the first test but did not reach ceiling for LF items until the end of the 2nd day (see Table 1). Eye position was coded for each frame on the videotape record beginning 500 ms before target onset (to ensure there were no display biases that favored the critical items prior to the spoken instruction) and ending when the participant clicked on a shape. The 1st day's test was lost for 3 participants because of equipment failures or experimenter error. The 2nd day's test was coded for all participants. In order not to overestimate competitor fixations, we coded only trials on which participants selected the correct object with a mouse click.

Table 1
Accuracy in Training and Testing in Experiment 1

Block	Overall	HF	LF
Training 1 (2AFC)	0.728	0.751	0.562
Training 4 (2AFC)	0.907	0.933	0.722
Training 7 (4AFC)	0.933	0.952	0.797
Day 1 test	0.863	0.949	0.777
Training 8 (4AFC)	0.940	0.960	0.802
Training 11 (4AFC)	0.952	0.965	0.859
Training 14 (4AFC)	0.969	0.977	0.908
Day 2 test	0.974	0.983	0.964

Note. HF = high frequency; LF = low frequency; 2AFC = two-alternative forced choice; 4AFC = four-alternative forced choice.

Figure 2 shows the proportion of fixations to cohort, rhyme, and unrelated distractors on Days 1 and 2 in 33-ms time frames (30 Hz, video sampling rate) from the onset of the name of the target in the "Click on the . . . " instruction. Proportions are averaged across all frequency and neighbor (cohort or rhyme) conditions for the test on Day 1 ($n = 13$) and Day 2 ($n = 16$). On both days cohorts and rhymes were fixated more than unrelated distractors. (Note that fixation probabilities for unrelated items represent the *average* fixation proportion to all unrelated items.) The overall pattern on Day 2 was strikingly similar to the pattern Allopenna et al. (1998) found with real words. The cohort and target proportions separated together from the unrelated baseline. Slightly later, the fixation proportions to the rhyme separated from baseline.

Eye movements were more closely time-locked to speech than it may appear on first inspection of Figure 2. The minimum latency to plan and launch a saccade in simple tasks is estimated to be between 150 and 180 ms (e.g., Fischer, 1992; Saslow, 1967), whereas typical intersaccadic intervals in tasks more comparable with ours (e.g., visual search) fall in the range of 200 to 300 ms (Viviani, 1990). Allowing for an estimated 200 ms to plan and

[6] The possible unrelated items included the "near neighbor"—the item differing from the target by two phonemes (the initial consonant and final vowel), which would not be predicted to compete under the one-phoneme difference neighborhood metric. Of the test trials with three unrelated items, 50% of these included the near neighbor.

[7] On 79% of the 2AFC training trials, the distractor was unrelated. The remaining 21% were evenly distributed into trials with cohort, rhyme, or near neighbor (the item differing in the first consonant and final vowel from the target; e.g., /pibo/ and /dibu/ are near neighbors) distractors. On 46% of the 4AFC training trials, all three items were unrelated. On 45% of the 4AFC training trials, the distractors were two unrelated items and one item from the target's four-word set (15% for each related type; if we count the near neighbor as an unrelated item, 61% of trials had three unrelated distractors). On the remaining 9% of trials, one unrelated and two related items were the distractors (cohort + rhyme, cohort + near neighbor, and rhyme + near neighbor each accounted for 3% of trials).

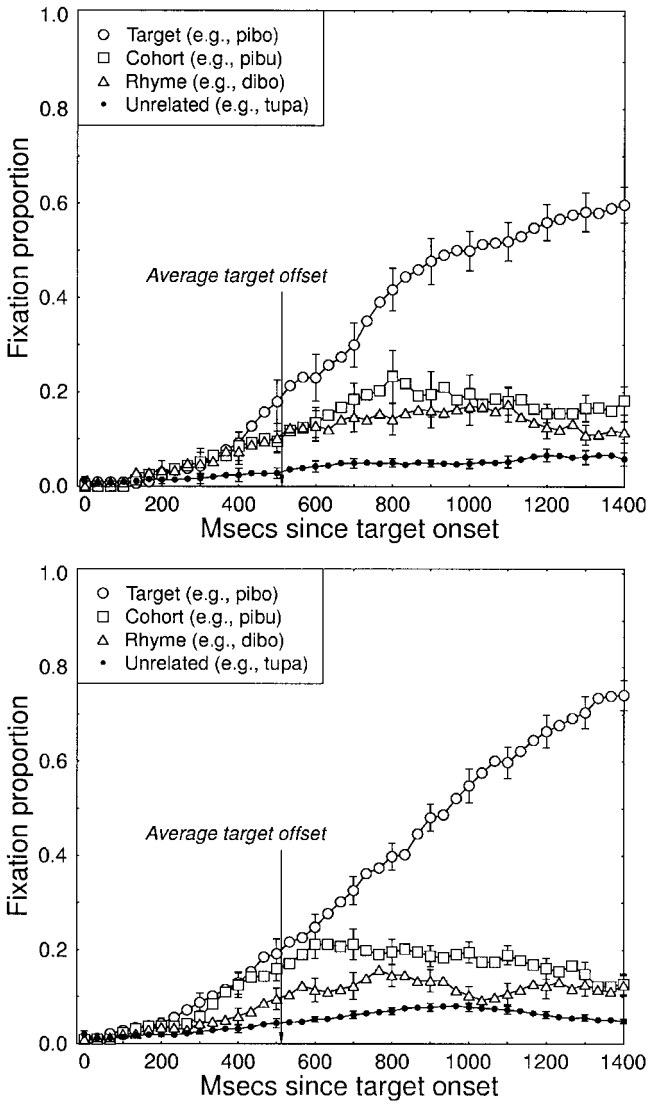

Figure 2. Fixation probabilities averaged over cohort and rhyme trials. Upper panel: Day 1 test; lower panel: Day 2 test. Standard error bars are shown for every third data point.

launch a saccade, the earliest eye movements were being planned almost immediately after target onset. Because the average target duration was 496 ms, eye movements in approximately the first 700 ms were planned prior to target offset.

Note that the slope of the target fixation proportion (derived from a logistic regression) was less than for real words (Day 1: proportion increased 0.0006/ms; Day 2: 0.0007; real words [Allopenna et al., 1998]: 0.0021), and the target proportion did not reach 1.0 even 1,500 ms after the onset of the target name. Two factors underlie this. First, the stimuli were longer than the bisyllabic words used by Allopenna et al. because of their CV-CV structure and unreduced vowels. Second, although participants were at ceiling on HF and LF items in the second test (Table 1), they were apparently not as confident as we would expect them to be with real words, as indicated by the fact that they made more eye movements than participants in Allopenna et al. (1998): 3.4/trial on Day 2 versus 1.5/trial for real words. Two differences stand out between the results for Days 1 and 2. First, the increased slope for target fixation probabilities on Day 2 likely reflects additional learning; participants were able to resolve competition among items more quickly as training progressed. Second, the rhyme effect on Day 1 appeared to be about as strong as the cohort effect, a point we discuss in detail in the "Simulations of Experiments 1 and 2" section below.

Cohort and rhyme competitor effects were modulated by target and neighbor frequency. This is illustrated in Figures 3 and 4 with data from the second day's test. Results were similar, but noisier, on Day 1.

To quantify the differences apparent in the figures, we conducted ANOVAs on mean fixation proportions[8] in the time window from 200 ms after target onset to 1,400 ms. We chose to begin this window at 200 ms because that is approximately the earliest point at which we expect to observe differences in fixation proportions due to reactions to the speech stimulus. The 1,400-ms endpoint corresponds to the time by which target fixation proportions asymptoted across participants.

We analyzed target and competitor fixation proportions for critical trials (i.e., those in which a cohort or rhyme was among the distractors). For target proportions, we used a 2 (day) $\times$ 2 (target frequency) $\times$ 2 (competitor frequency) $\times$ 2 (competitor type: cohort or rhyme) analysis. The predictions were that (a) HF targets should have higher fixation proportions than LF targets, (b) targets presented with HF competitors should be fixated less than those presented with LF competitors (i.e., if those distractors compete for recognition with strength proportional to their frequency), and (c) given previous results with real words, cohorts should compete more than rhymes, as would be indicated by lower target fixation proportions for cohort than rhyme trials.

We conducted an analysis of the fixation proportions to cohorts, rhymes, and unrelated distractors using the same structure, except that there were three levels of competitor type (because we included fixation proportions to unrelated items). The predictions for this analysis mirrored those for the target analysis: There should be lower fixation proportions for competitors displayed with HF targets, higher proportions for HF competitors, and higher proportions for cohorts than rhymes (and higher proportions for rhymes than unrelated items).

We did not have a priori expectations of interactions among these factors in either analysis, although the time course plots suggest we might find a Competitor Type $\times$ Day interaction, because the magnitude of the rhyme effect appeared to decrease in the 2nd day (see Figure 2). We did expect to find stronger effects in general after 2 days of training, because the effect of training should be to strengthen the representations of the artificial lexical items in memory. Given this expectation (bolstered by the fact that participants did not reach ceiling levels of accuracy on LF items

[8] A number of strategies have been used for analyzing the time course data provided by eye movements, such as multiple repeated measures ANOVAs on successive time windows. All strategies used, including this one, violate to some degree the assumptions underlying ANOVA. The current method of summing fixation proportions over a large time window minimizes the number of assumptions violated as well as the degree.

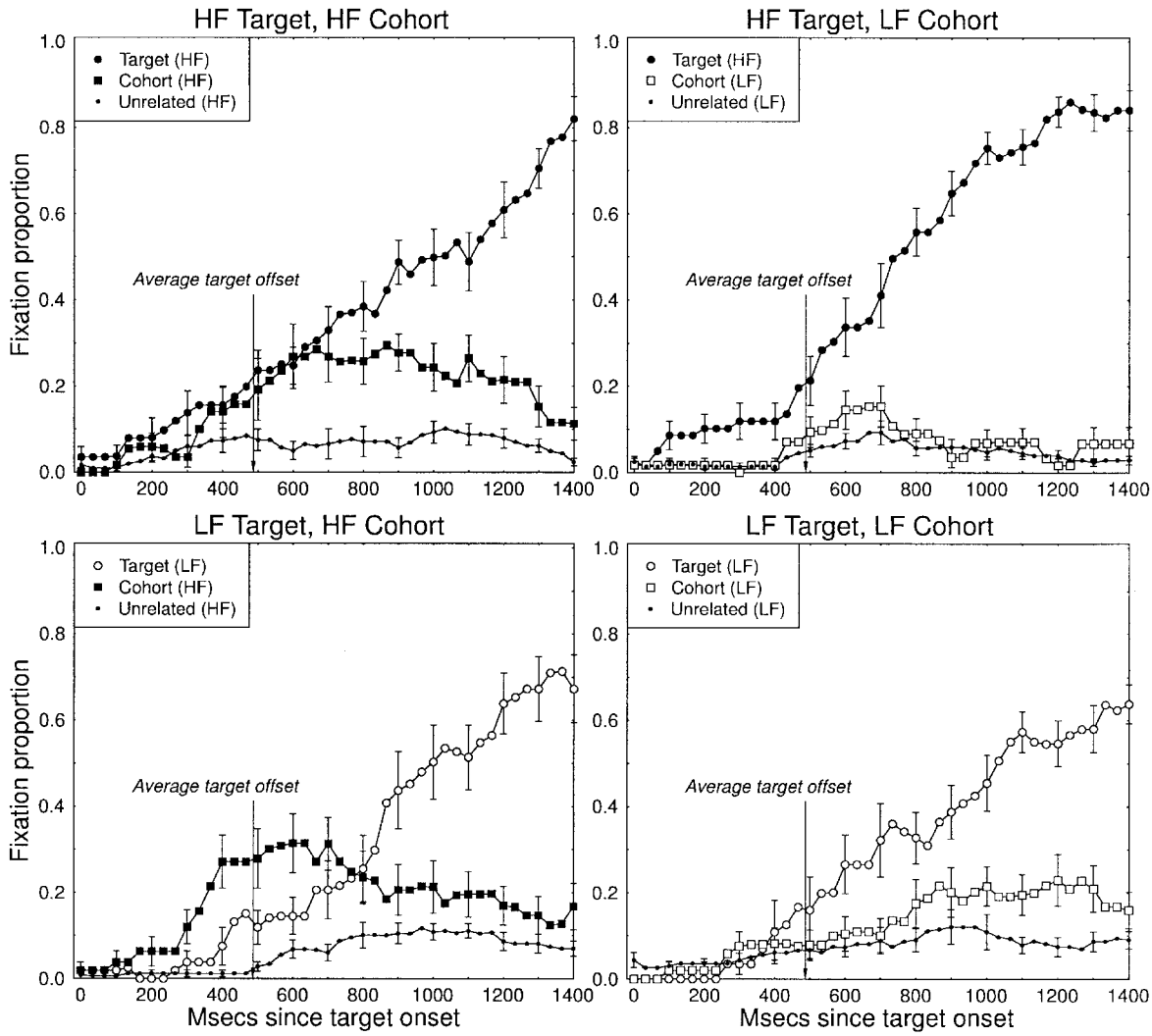

Figure 3. Cohort effects as modulated by target and neighbor frequency on the Day 2 test of Experiment 1. High-frequency (HF) targets are shown in the upper panels, and low-frequency (LF) targets are shown in the lower panels. Targets with HF competitors are shown in the two left panels, and targets with LF competitors are shown in the two right panels. Standard error bars are shown for every third data point.

until Day 2) and that we had data from 3 more participants available for the Day 2 analysis, we also conducted planned analyses for each day of training.

Target fixation proportions. We begin with the results from the analyses of target fixation proportions. The ANOVA results[9] for the full analysis are shown in Table 2. We focus on the main effects and the one significant interaction (Target × Competitor Frequency). The main effect of day was not significant (target fixations increased slightly, from .35 [*SD* = .21] on Day 1 to .40 [*SD* = .17] on Day 2), nor did day interact reliably with any other factor. The same was true for competitor type; overall, targets presented with cohorts and rhymes received nearly identical fixation proportions (with cohorts = .37, *SD* = .19, with rhymes = .38, *SD* = .19). There was a strong main effect of target frequency, with a higher fixation proportion for HF targets (.43, *SD* = .17) than LF targets (.32, *SD* = .19). There was a strong trend toward an effect of competitor frequency, with a lower fixation proportion for targets presented with HF competitors (.35, *SD* = .18) than with LF competitors (.40, *SD* = .20). Target and competitor frequency interacted reliably. An examination of simple effects showed there were reliable effects of target frequency for both levels of competitor frequency (Fs > 4.9, ps < .05, ω^2s > .13).

[9] We estimated effect size using partial ω^2 as described in Keppel (1991) for within-subjects designs. The goal of the measure is to estimate the proportion of observed variance due to the treatment, or the amount of variability explained by the treatment. This is done by dividing total variance (treatment variance plus variance due to error) by treatment variance. Note that a common heuristic for interpreting this measure comes from Cohen (1977), who suggested values greater than .15 be considered large, values greater than .06 medium, and values greater than .01 small.

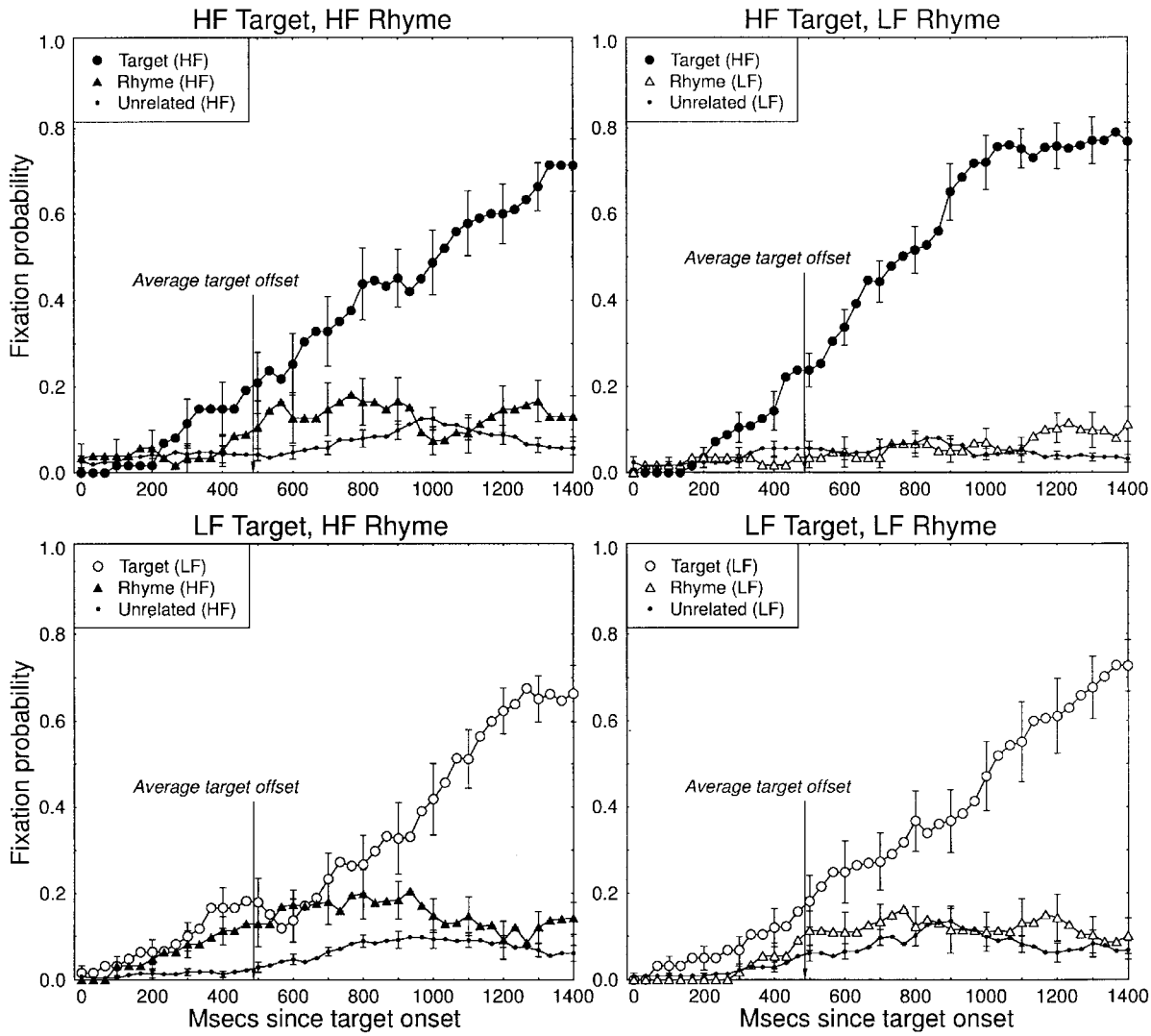

Figure 4. Rhyme effects as modulated by target and neighbor frequency on Day 2 of Experiment 1. High-frequency (HF) targets are shown in the upper panels, and low-frequency (LF) targets are shown in the lower panels. Targets with HF competitors are shown in the two left panels, and targets with LF competitors are shown in the two right panels. Standard error bars are shown for every third data point.

The interaction was due to a strong effect of competitor frequency for HF targets, $F(1, 12) = 7.0$, $p = .02$, $\omega^2 = .19$ (HF competitor: $M = .39$, $SD = .14$; LF competitor: $M = .48$, $SD = .10$), but virtually no effect on LF targets, $F(1, 12) = 0.2$, $p = .68$ (HF competitor: $M = .31$, $SD = .10$; LF competitor: $M = .32$, $SD = .15$). This suggests that although the representations of LF targets were sufficiently strong to support high accuracy, they were weak enough that the presence of an HF competitor was not sufficient to slow recognition detectably compared with the LF competitor condition.

Although there were no effects of day in the full analysis, the planned separate analyses by day revealed substantial differences in the effect of competitor frequency. The analyses are summarized in Tables 3 and 4. In the analysis of Day 1 ($n = 13$), the only reliable effect was that of target frequency (HF = .42, $SD = .18$; LF = .28, $SD = .21$), $F(1, 12) = 15.6$, $p = .002$, $\omega^2 = .36$. There was an unreliable trend for lower fixation proportions for targets presented with HF competitors (.33, $SD = .19$) than LF competitors (.37, $SD = .23$).

In the Day 2 analysis with all 16 participants, target frequency was the only effect reliable at the $\alpha = .05$ level (HF = .46, $SD = .15$; LF = .35, $SD = .17$), $F(1, 12) = 14.6$, $p = .002$, $\omega^2 = .30$, but there were also two strong trends reliable at the $\alpha = .10$ level. Targets presented with HF competitors were fixated less than those presented with LF competitors (HF = .37, $SD = .17$; LF = .43, $SD = .16$), and there was a trend toward Target × Competitor Frequency interaction. Simple effects analyses showed that the pattern was similar to that found in the overall analysis; there was a strong effect of competitor frequency on HF targets (HF = .40, $SD = .13$; LF = .51, $SD = .09$), $F(1, 15) = 8.9$, $p = .009$, $\omega^2 = .20$, but not on LF targets (HF = .34, $SD = .15$, LF = .36, $SD = .15$), $F(1, 15) = 0.1$, $p = .71$, $\omega^2 < 0$. In addition, there was a

Table 2
Analysis of Variance for Target Fixation Proportions (N = 13)

Source	*F*	ω^2	*p*
Day (D)	2.97	.07	.111
Target frequency (TF)	20.02**	.42	.001
D × TF	2.50	.05	.140
Competitor frequency (CF)	3.39†	.08	.091
D × CF	0.54	0	.478
TF × CF	5.14*	.14	.043
D × TF × CF	0.50	0	.492
Competitor type (CT)	0.16	0	.695
D × CT	0.22	0	.645
TF × CT	0.68	0	.426
D × TF × CT	0.35	0	.567
CF × CT	1.32	.01	.272
D × CF × CT	0.08	0	.777
TF × CF × CT	0.01	0	.911
D × TF × CF × CT	2.40	.05	.148

Note. Degrees of freedom for each comparison are 1, 12. Competitor type was either cohort or rhyme.
† $p < .10$. * $p < .05$. ** $p < .01$.

strong effect of target frequency on LF competitors, $F(1, 16) = 37.6$, $p < .001$, $\omega^2 = .54$, but not on HF competitors, $F(1, 16) = 1.7$, $p = .218$, $\omega^2 = .02$. The separate analyses for each day show that whereas the effect of target frequency was apparent early in learning, the effect of competitor frequency emerged more slowly.

Competitor fixation proportions. We turn now to the analysis of fixation proportions to potential competitors (cohorts, rhymes, and unrelated distractors). The full analysis including the day factor is summarized in Table 5. As with the target proportion analysis, the main effect of day was not significant. There was a main effect of target frequency: Competitors displayed with HF targets were fixated less ($M = .09$, $SD = .09$) than those presented with LF targets ($M = .11$, $SD = .10$). The main effect of competitor frequency was also reliable: HF competitors were fixated more ($M = .12$, $SD = .11$) than LF competitors ($M = .08$, $SD = .08$). There was a reliable Day × Competitor Frequency interaction. We determine the basis of this interaction in the separate analyses by day below. There was a strong main effect of competitor type. Planned comparisons showed that fixation proportions were reliably greater to cohorts ($M = .13$, $SD = .06$) than unrelated items

Table 3
Analysis of Variance for Target Fixation Proportions for Day 1 (N = 13) in Experiment 1

Source	*F*	ω^2	*p*
Target frequency (TF)	15.57**	.36	.002
Competitor frequency (CF)	1.44	.02	.254
TF × CF	1.40	.01	.261
Competitor type (CT)	0.51	0	.489
TF × CT	0.65	0	.435
CF × CT	1.24	0	.309
TF × CF × CT	0.90	0	.361

Note. Degrees of freedom for each comparison are 1, 12. Competitor type was either cohort or rhyme.
** $p < .01$.

Table 4
Analysis of Variance for Target Fixation Proportions for Day 2 (N = 16) in Experiment 1

Source	*F*	ω^2	*p*
Target frequency (TF)	14.60**	.30	.002
Competitor frequency (CF)	3.48†	.07	.082
TF × CF	3.29†	.07	.090
Competitor type (CT)	0	0	.97
TF × CT	.23	0	.639
CF × CT	.02	0	.883
TF × CF × CT	.13	.02	.728

Note. Degrees of freedom for each comparison are 1, 16. Competitor type was either cohort or rhyme.
† $p < .10$. ** $p < .01$.

($M = .06$, $SD = .03$), $F(1, 12) = 30.1$, $p < .001$, $\omega^2 = .53$, and rhymes ($M = .11$, $SD = .05$) were fixated reliably more than unrelated items ($M = .06$, $SD = .03$), $F(1, 12) = 21.6$, $p = .001$, $\omega^2 = .44$. Cohorts were not fixated reliably more than rhymes, $F(1, 12) = 2.8$, $p = .121$, $\omega^2 = .06$. The only other effect to approach reliability was the Competitor Frequency × Competitor Type interaction (at $\alpha = .10$, $p = .082$). The simple effect of competitor frequency was reliable for cohort trials (HF = .17, $SD = .07$; LF = .10, $SD = .06$), $F(1, 12) = 7.3$, $p = .019$, $\omega^2 = .20$; it approached reliability for rhyme trials (HF = .13, $SD = .06$; LF = .10, $SD = .05$), $F(1, 12) = 2.5$, $p = .13$, $\omega^2 = .06$; and there was no effect on unrelated items (HF = .06, $SD = .03$; LF = .06, $SD = .02$), $F(1, 12) = .40$, $p = .54$, $\omega^2 = 0$. The absence of an effect on unrelated items demonstrates that frequency alone did not have much influence on the likelihood of fixation; rather, it modulated the likelihood of fixations to items with substantial phonological similarity to the target.

The results of the planned separate analyses of competitor fixations by day are shown in Tables 6 and 7, and as for the target analyses, there were substantial differences between the 2 days. On Day 1 (see Table 6, $n = 13$), the only reliable effect was of

Table 5
Analysis of Variance for Fixation Proportions to Competitors in Experiment 1

Source	*df*	*F*	ω^2	*p*
Day (D)	1, 12	0	0	.997
Target frequency (TF)	1, 12	7.17*	.19	.020
D × TF	1, 12	0.07	0	.797
Competitor frequency (CF)	1, 12	7.70*	.20	.017
D × CF	1, 12	6.80*	.18	.023
TF × CF	1, 12	1.03	0	.331
D × TF × CF	1, 12	0.46	0	.513
Competitor type (CT)	2, 24	20.85**	.50	< .001
D × CT	2, 24	1.64	.03	.214
TF × CT	2, 24	0.51	0	.608
D × TF × CT	2, 24	0.10	0	.909
CF × CT	2, 24	2.78†	.08	.082
D × CF × CT	2, 24	1.88	.04	.174
TF × CF × CT	2, 24	1.41	.02	.263
D × TF × CF × CT	2, 24	0.15	0	.863

Note. Competitor types were cohort, rhyme, and unrelated.
† $p < .10$. * $p < .05$. ** $p < .01$.

Table 6
Analysis of Variance for Fixation Proportions to Competitors for Day 1 (n = 13) in Experiment 1

Source	*df*	*F*	ω^2	*p*
Target frequency (TF)	1, 12	2.63	.06	.131
Competitor frequency (CF)	1, 12	1.08	0	.319
TF × CF	1, 12	0.00	0	.972
Competitor type (CT)	2, 24	10.38**	.27	.001
TF × CT	2, 24	0.35	0	.705
CF × CT	2, 24	0.32	0	.733
TF × CF × CT	2, 24	0.31	0	.739

Note. Competitor types were cohort, rhyme, and unrelated.
** $p < .01$.

competitor type. The patterns of reliability were the same as in the overall analysis: cohorts ($M = .13$, $SD = .07$) and rhymes ($M = .13$, $SD = .06$) were fixated reliably more than unrelated items ($M = .06$, $SD = .03$; Fs > 11.6, ps $< .005$, ω^2s $> .28$), but the cohort fixation proportion was not greater than that of rhymes.

On Day 2, the pattern was much different. There was a strong trend toward an effect of target frequency (HF = .09, $SD = .09$; LF = .12, $SD = .10$), as well as a significant effect of competitor frequency (HF = .13, $SD = .11$; LF = .08, $SD = .07$). The Target × Competitor Frequency interaction was nearly reliable as well. Analyses of simple effects showed this to be due to reliable effects of competitor frequency at both levels of target frequency, Fs(1, 15) > 7.5, ps $< .05$, ω^2s $> .16$, but whereas there was a reliable effect of target frequency on LF competitors (HF = .06, $SD = .03$; LF = .11, $SD = .06$), $F(1, 15) = 9.5$, $p = .008$, $\omega^2 = .21$, there was virtually no effect of target frequency on HF competitors (HF = .13, $SD = .05$; LF = .14, $SD = .06$), $F(1, 15) = .20$, $p = .663$, $\omega^2 = 0$. This is a complementary pattern to that found in the simple effects analyses of the Target × Competitor Frequency interaction in the Day 2 analysis of target fixation proportions: Effects of competitor frequency were most apparent on trials with HF targets (reducing target proportions and increasing competitor proportions). Although there were trends in the predicted directions for effects of competitor frequency on target and competitor proportions in trials with LF targets, the effects were much smaller. In other words, we observed nearly as much competition from LF competitors as from HF competitors on LF targets. This suggests, as we discussed in the full analysis of target fixation proportions, that although the representations of LF targets participants acquired in training were strong enough to support high accuracy, they were weak enough that we observed a floor effect in terms of the influence of competitor frequency on these targets.

In summary, HF targets were fixated more than LF targets, HF competitors were fixated more than LF competitors, and targets presented with LF competitors were fixated more than those presented with HF competitors. Thus, our training paradigm was successful in instantiating levels of frequency roughly analogous to those found with natural language materials: HF items were more easily recognized than LF items and competed more strongly when they were potential competitors. We observed reliably more fixations to cohort and rhyme competitors than to unrelated items in both tests. The magnitude of the cohort effect grew with training, whereas the rhyme effect weakened (as can be seen in Figure 2). Thus, training led to reliable competition effects similar to those found with real words after a single training session, although the magnitude of rhyme competition relative to that of cohort competition was larger than has been observed with real words (e.g., Allopenna et al., 1998). With more training, the relative magnitude of cohort and rhyme effects came to resemble more closely the pattern observed with real words (i.e., more and earlier fixations to cohorts than rhymes).

Discussion

With relatively little training (98 exposures to HF items and 14 exposures to LF items), the time course of processing novel words closely resembled that of real words. In fact, after just 49 exposures to HF items and 7 exposures to LF items on the 1st day of training, cohort and rhyme effects were already present. These results from an artificial lexicon replicate previous results found with real words, including cohort and rhyme competition (Allopenna et al., 1998) and the time course of frequency effects (Dahan, Magnuson, & Tanenhaus, 2001). Moreover, the results demonstrate that the artificial lexicon paradigm can be used effectively to study the processing of newly learned lexical items; with just a few hours of training, the artificial lexical items provided sufficient analogs to real words to replicate previous results and to provide the first time course measures of neighborhood density effects. In addition, two new results emerged. First, item frequency modulated competitor effects for rhymes just as it did for cohorts. This result strongly supports models such as the neighborhood activation model (P. A. Luce & Pisoni, 1998) and TRACE (McClelland & Elman, 1986), in which an initial mismatch does not inhibit a potential lexical candidate so strongly as to eliminate it from the activation set. Second, the pattern of rhyme and cohort competition changed with learning. Initially, both rhyme and cohort competitors received an equivalent proportion of fixations and showed a similar time course. With more exposure, however, the rhyme effects became somewhat diminished, and they showed delayed onset compared with the cohorts. This result is explored with a computational model after we describe Experiment 2. Although Experiment 1 arguably demonstrates direct and indirect effects of lexical competition (in competitor and target frequency effects), it leaves open the possibility that these effects are due to the characteristics of targets and competitors displayed simultaneously. Experiment 2 examines whether the eye-tracking para-

Table 7
Analysis of Variance for Fixation Proportions to Competitors for Day 2 (N = 16) in Experiment 1

Source	*df*	*F*	ω^2	*p*
Target frequency (TF)	1, 15	4.12†	.09	.061
Competitor frequency (CF)	1, 15	27.90**	.46	< .001
TF × CF	1, 15	3.68†	.08	.074
Competitor type (CT)	2, 30	20.37**	.38	< .001
TF × CT	2, 30	0.44	0	.651
CF × CT	2, 30	7.09**	.16	.003
TF × CF × CT	2, 30	0.38	0	.684

Note. Competitor types were cohort, rhyme, and unrelated.
† $p < .10$. ** $p < .01$.

digm is sensitive to lexical competition between items that are not simultaneously displayed.

Experiment 2

The eye-tracking paradigm differs in important ways from conventional psycholinguistic measures: It provides a fine-grained, on-line measure of lexical processing in continuous speech, it allows us to monitor the activation of multiple lexical items simultaneously, and it uses a naturalistic task—following a spoken instruction rather than making metalinguistic responses to or repeating typically decontextualized stimuli. However, these last two points in particular might also be viewed as limitations of the paradigm, as the use of visual displays raises at least two concerns. First, the paradigm might not be sensitive to effects of nondisplayed lexical competitors (which other methods, such as identification in noise or lexical decision, are; P. A. Luce & Pisoni, 1998). This would make it difficult to examine effects of lexical neighborhood density except by displaying all of a word's neighbors. Second, the observed effects might depend crucially on interactions between pictured referents and their names rather than primarily reflecting input-driven lexical activation.

Experiment 2 examines whether the neighborhood density effects observed in Experiment 1 depend on the display of pictures of potential competitors. Experiment 2 asked the following question: Does the *frequency* of an item's neighbors slow the time course of recognition (as it does in tasks such as identification in noise; e.g., P. A. Luce & Pisoni, 1998) even when the neighbors are not displayed? This result would complement that of Dahan, Magnuson, Tanenhaus, and Hogan (2001), who showed that subtle alterations of words (misleading coarticulatory cues generated by cross-splicing two CVC words, such as *neck* and *net*, near the offset of the vowel) that made the target word temporarily consistent with another word in the lexicon—a word that was neither displayed nor mentioned—delayed looks to the pictured referent. For the present study, we included the cohort, rhyme, and frequency conditions from Experiment 1. In addition, we compared the time course of recognition for HF and LF words with HF and LF neighbors when the neighbors were *not* displayed. If neighborhood characteristics influence the rise time of fixation probabilities when those neighbors are not displayed, this will demonstrate that fixation probabilities reflect competition within the entire artificial lexicon rather than just properties of the displayed alternatives.

Method

Participants. Eight students at the University of Rochester were paid $7.50/hr for their participation. All were native speakers of English with normal hearing and normal or corrected-to-normal vision.

Materials and procedure. Experiment 2 differed from Experiment 1 in that a third level of frequency was used. Half of the items were presented with medium frequency (MF). Six items were HF, two were LF, and eight were MF. All of the MF items had MF neighbors. The HF and LF items were assigned such that four of the HF items had HF neighbors, and two had LF neighbors (and the neighbors for the two LF items were those two HF items). The MF level provided a set of items that could be used as distractors without changing the ratio of HF to LF items.

Each training block consisted of 68 trials. HF items appeared seven times per block, LF items appeared once per block, and MF items appeared three times per training block. The tests consisted of 96 trials. Each item appeared in 6 trials: one with its cohort (onset) neighbor and two unrelated items, one with its rhyme (offset) neighbor and two unrelated items, and four with three unrelated items. For the crucial comparisons (HF targets having HF or LF neighbors but displayed with three unrelated distractors), MF items were used as unrelated distractors so that any difference in target probabilities cannot be attributed to distractor characteristics.

Results

Participants reached ceiling levels of accuracy by the end of Day 2 (see Table 8). Experiment 2 replicated the basic cohort and rhyme patterns found in Experiment 1, as well as the same pattern of frequency effects. For sake of brevity, we will not examine these results in detail, but instead focus on the influence of absent competitors. Figure 5 shows the results of the crucial conditions after the 2nd day of training: the fixation probabilities for HF targets with HF or LF neighbors presented among unrelated, MF distractors. As predicted, the fixation probabilities for targets with LF neighbors rose more quickly than for targets with HF neighbors.

We conducted 2 (day) $\times$ 2 (competitor frequency) ANOVAs on mean target fixation proportions and mean unrelated distractor fixation proportions in the same 200–1,400 ms window as in Experiment 1 for the HF target condition. Again, the prediction is that we should observe reliably higher fixation proportions when a target's competitors are LF than when they are HF—even when they are not displayed. The analysis of unrelated distractor proportions provides a test of an alternative basis for differences in target proportions. Although we predict that targets with HF competitors should be fixated less due to lexical competition, the differences could result from competition with the three unrelated distractors. Although there should be no basis for a differential effect because all the unrelated distractors were of MF, this analysis will verify that we do not observe commensurate increases in unrelated fixation proportions when we observe decreases in target proportions, and vice versa. As in Experiment 1, we also planned separate analyses of fixations from each day's test, because we were primarily interested in the results once training was complete.

In the overall ANOVA on target proportions, we found a reliable main effect of day, as the proportion increased from .38 (SD = .20) on Day 1 to .50 (SD = .17) on Day 2, $F(1, 7) = 7.9$, $p = .026$, $\omega^2 = .30$. The main effect of competitor frequency was nearly reliable (HF = .39, SD = .16; LF = .48, SD = .21), $F(1,$

Table 8
Accuracy in Training and Testing in Experiment 2

Block	Overall	HF	MF	LF
Training 1 (2AFC)	0.680	0.738	0.594	0.500
Training 4 (2AFC)	0.948	0.969	0.917	0.857
Training 7 (4AFC)	0.912	0.943	0.902	0.625
Day 1 test	0.884	0.896	0.914	0.798
Training 8 (4AFC)	0.928	0.955	0.900	0.778
Training 11 (4AFC)	0.965	0.982	0.909	1.000
Training 14 (4AFC)	0.969	0.973	0.906	0.875
Day 2 test	0.962	0.966	0.925	0.933

Note. HF = high frequency; MF = medium frequency; LF = low frequency; 2AFC = two-alternative forced choice; 4AFC = four-alternative forced choice.

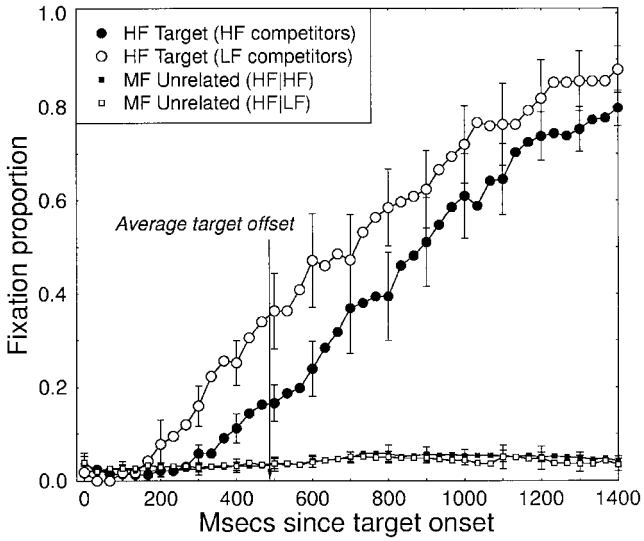

Figure 5. Fixation probabilities to high-frequency (HF) targets with HF or low-frequency (LF) neighbors when presented among three unrelated, medium-frequency (MF) distractors on the Day 2 test in Experiment 2. Standard error bars are shown for every third data point.

7) = 4.4, $p = .074$, $\omega^2 = .18$. Although the Day × Competitor Frequency interaction was not reliable, $F(1, 7) = 1.0$, $p = .349$, the effect of competitor frequency differed substantially in the planned separate analyses of each day. On Day 1, the effect was not reliable (HF = .36, *SD* = .17; LF = .40, *SD* = .22), $F(1, 7) = 0.4$, $p = .558$. On Day 2, however, there was a strong effect of competitor frequency (HF = .43, *SD* = .14; LF = .56, *SD* = .17), $F(1, 7) = 9.4$, $p = .018$, $\omega^2 = .34$. The lack of a Day × Competitor Frequency interaction in the overall analysis stems from similar effects of competitor frequency at each level of day and similar effects of day at each level of competitor frequency, but this masks the fact that the frequency effect became stronger with additional training.

The analyses of unrelated distractor fixations showed that competition among displayed items cannot account for differences in target proportions as a function of competitor frequency, because we did not observe commensurate increases in unrelated fixation proportions for decreases in target proportions or vice versa. In the overall analysis and analyses by day, in both the 200–1,400 ms and 200–1,000 ms time windows, there were not significant effects of day, competitor frequency, or any significant interactions (*F*s < 2.5, *p*s > .15). Mean unrelated fixation proportion decreased from .07 (*SD* = .07) to .04 (*SD* = .03) from Day 1 to Day 2, but proportions never differed by more than .02 as a function of competitor frequency.

Discussion

The results of Experiment 2 show that the eye-movement paradigm reveals lexical processing that extends beyond those items present in the visual display: The time course of recognition depended on characteristics of *nondisplayed* neighbors. This is an important demonstration not just for the artificial lexicon paradigm but also for eye tracking during spoken language processing more generally. Although we explicitly assume that the context of the visual display affects fixation behavior through the set of alternatives available for response selection (Allopenna et al., 1998; Dahan, Magnuson, & Tanenhaus, 2001; Dahan, Magnuson, Tanenhaus, & Hogan, 2001), Experiment 2 shows that the lexical processing on which fixations are based is not constrained to the displayed set.

Simulations of Experiments 1 and 2

The cohort and rhyme effects as well as the frequency effects observed in Experiments 1 and 2 are similar to those we have observed with English lexical items, which we have modeled successfully using TRACE (McClelland & Elman, 1986) and a simple linking hypothesis relating lexical activation to fixation probability (Allopenna et al., 1998; Dahan, Magnuson, & Tanenhaus, 2001). However, because TRACE and most other current word recognition models (e.g., Shortlist [Norris, 1994] and PARSYN [P. A. Luce, Goldinger, Auer, & Vitevitch, 2000]) are not learning models, they cannot account for a key aspect of our results: the change in the relative and absolute magnitudes of rhyme and cohort effects between the 1st and 2nd days of training. We decided to use SRNs (Elman, 1990) to model these effects because SRNs are particularly well suited to the processing of sequential input, and relatively small networks can capture temporal dependencies (Dell, Juliano, & Govindjee, 1993; Elman, 1990; Jordan, 1986; Norris, 1990).

The present simulations had two goals. The first goal was to determine whether a simple learning model would capture the posttraining patterns in the data—effects we have previously modeled using TRACE. The second goal was to determine whether a learning-based model would capture the changes during training, in particular the decreased proportion of looks to rhyme competitors on Day 2 compared with Day 1.

Method

Network. The architecture of the network is presented in Figure 6. The input layer consisted of 18 units, corresponding to the following phonetic features: syllabic, consonantal, sonorant, nasal, anterior, coronal, continuant, delayed release, strident, voicing, aspiration, lateral, high, low, back,

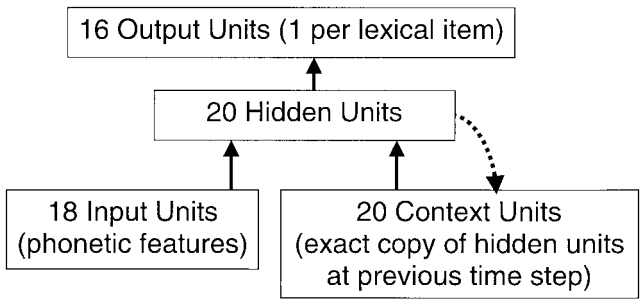

Figure 6. Schematic of the networks used. Solid arrows indicate fully connected (e.g., every input unit has a connection to every hidden unit), trainable links. The dotted arrow indicates one-to-one copy back connections (the first hidden unit connects only to the first context unit, the second hidden unit connects only to the second context unit, etc.) with fixed weights (1.0) and a one-step time delay.

round, tense, and reduced. Thus, the bottom-up input was a distributed representation based on basic phonetic principles. Each input unit had a feed-forward connection to each of 20 hidden units (18 × 20 connections). Each of these connections had a modifiable connection weight. Each hidden unit had a feed-forward, trainable connection to each of 16 output units (a localist representation of the 16-word artificial lexicon; thus, 20 × 16 connections). Each hidden unit also had a one-to-one feed-forward, fixed-weight (1.0) connection to a corresponding context unit (20 × 1 connections). Thus, at Time t +1, the state of the third context unit, for example, would become identical to the state of the third hidden unit at Time t. These context units had trainable forward connections to each hidden unit (20 × 20 connections). As is standard in such models, a single bias unit (with constant activation of 1.0) was fully connected to the hidden and output layers.

Materials and procedure. Phonemes were represented as phonetic-feature vectors that were ramped on and off over several input cycles. To achieve a coarse analog of coarticulation similar to that used in TRACE, we overlapped phonemes, with a new phoneme beginning at the time step before the center of the preceding phoneme. Overlapping vectors were summed. To approximate differences in vocalic and consonantal prominence and duration, vowels were slightly longer than consonants. Figure 7 gives an example of the input corresponding to /pibo/, and details regarding the representational system are given in the Appendix.

Half the words were categorized as HF, following the scheme described for Experiment 1 (such that, e.g., /pibo/ and /dibu/ might be HF, whereas /dibo/ and /pibu/ might be LF, or all four might be HF or LF). Initially, all connections in the network were set to small, random values (with the exception of the hidden-to-context connections, which had constant weights of 1.0). For each training epoch, a list with three occurrences of each of the 8 low-frequency items and four occurrences of each of the 8 HF items was randomized (see the Appendix for details on the decision to use this 4:3 HF:LF ratio). The resulting 56-item list was randomized, and each word was presented one phoneme vector at a time over 17 time steps. Uniform random noise was added to the input units both during training and testing (see the Appendix). Adding noise moved our idealized speech inputs a small step closer to real speech conditions, in which the input is characterized by noise and variability, and slowed the time course of learning. Without noise, we obtain similar results (i.e., we account for all the same training and posttraining effects), but the network quickly jumps from strong rhyme effects early in training to minimal rhyme effects. So an advantage of adding noise is that it slows the progress of learning such that developmental trends during training are more easily observed.

Activation spread through the network as follows. First, the feature vector was applied to the input units. The input vector was multiplied by the 18 × 20 matrix of input-to-hidden connection weights to provide the bottom-up input to the network. The vector corresponding to the activations of the context units (0 at the very beginning of training) was multiplied by the 20 × 20 matrix of context-to-hidden connection weights. The total input to the hidden units was the sum of the bottom-up (phonetic features) and top-down (context) inputs (transformed by means of a logistic activation function). The resulting hidden-unit activation vector was then multiplied by the 20 × 16 hidden-to-output connection weight matrix and transformed with a logistic activation function to result in the 20-unit output vector. At each time step, the task of the network was to activate the output unit corresponding to the current word (and not activate the other word units). The actual output was compared with the desired output, error was computed, and during training, the error was passed through the modifiable connection weights in the network using standard back-propagation (see Elman, 1990, for a description with SRNs), with each unit being modified according to an estimate of its contribution to overall error. The learning rate was set to 0.01. After every fifth epoch, the learning rate was set to zero and the network was tested on the list of 16 lexical items. Following the test, the learning rate was returned to 0.01 and training continued. Twenty identical networks were each trained for 1,800 epochs.

Predictions

In Experiment 1, rhyme effects were not reliably different from cohort effects on Day 1 (top panel of Figure 2), but by the end of Day 2 (bottom panel of Figure 2), the cohort effect was earlier and stronger, as Allopenna et al. (1998) found with real words. This

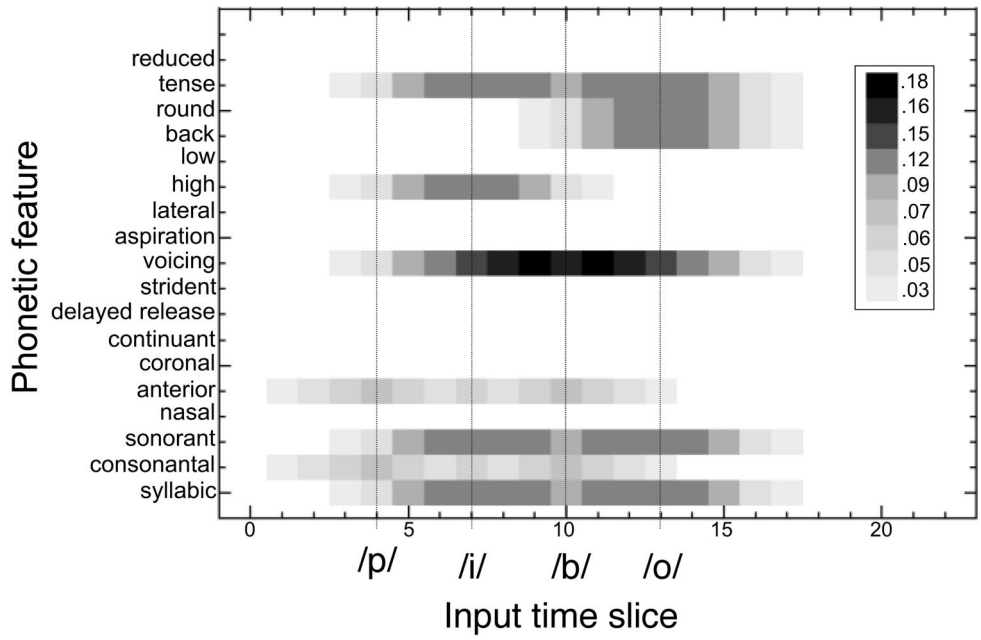

Figure 7. An example of the input used in the simple recurrent network simulations. Darkness indicates intensity. See text for details.

change is consistent with claims in the developmental literature that children's early lexical representations are more holistic than adults'. Some researchers have suggested that the structure of children's lexicons would allow holistic representations to distinguish between most words young children know (e.g., Charles-Luce & Luce, 1990). However, other analyses of child lexicons suggest fine-grained phonetic sensitivity is required (Coady & Aslin, in press; Dollaghan, 1994; but see Charles-Luce & Luce, 1995), and there is growing evidence that even very young children are sensitive to fine-grained, sequential phonetic information (Gerken, Murphy, & Aslin, 1995; Swingley & Aslin, 2000, 2002; Swingley, Pinto, & Fernald, 1999).

The change in relative cohort and rhyme effect magnitudes observed in our adult participants could be interpreted as evidence of initially holistic processing. On Day 1, they seemed not to be terribly sensitive to the greater onset overlap of cohorts and targets, because we observed approximately equivalent cohort and rhyme competition. There are at least two possible bases for the shift from equivalent competition effects to the cohort advantage on Day 2. First, our adult participants may be recapitulating the hypothetical holistic-to-incremental shift some researchers claim occurs in the development of spoken word processing. The fact that cohorts and rhymes seem to be considered equally good candidates for a target pattern suggests that, as is proposed for young children, the adults' initial representations of these new words are not incrementally specific; for example, the representations may be based more on overall similarity between syllables than precisely ordered segmental representations. Second, processing may be incremental from the beginning of learning, but early on, overall similarity may have greater impact than onset similarity because representations are weak early in learning. That is, the model has not had enough training to establish the reliability of its input and does not "commit" to a lexical hypothesis based on word onset. As learning progresses, the model "discovers" the relative diagnosticity of onset information. As a further illustration, consider the TRACE model. In TRACE, the basis for an advantage of onset similarity is that items overlapping with an input at onset become sufficiently active that they strongly inhibit items with greater overall similarity but that mismatch at onset. Given initially weak lexical representations as new words are learned (e.g., low weights between phoneme and lexical nodes in TRACE), the onset similarity advantage would be weakened; indeed, given weak representations, we might expect a reversal, with an advantage for overall similarity.

The SRN results would be consistent with the first explanation (holistic to incremental shift) if we do not see large differences in the time course of activation of cohorts and rhymes. In our stimuli, both competitor types overlap with targets by the same number of phonemes. On the holistic account, there is not a basis for expecting one or the other competitor type to have an advantage early in learning. On the second explanation (low confidence for weak representations), we would expect to see important differences in time course even early in learning. If processing is incremental, and the change in competitor effect magnitude is a result of strengthening representations, even early in learning we should observe earlier cohort activation than rhyme activation, with a rhyme advantage appearing late in the trial-by-trial time course.

Results

Mean error per test (the summed absolute differences between the states of the output units and the desired states) started out around .095. This was because the initial small, random weights only weakly activated the word units (i.e., initially all word units had activations around 0, making mean error low because the desired response would be 0 for all but one unit). Error quickly increased, and then began to decrease substantially after about 600 epochs. It then varied between .009 and .126, with the mean shifting from .09 between Epochs 600 and 900 to .03 between Epochs 1,500 to 1,800.

To facilitate a qualitative comparison with the eye movement data from Experiments 1 and 2, we computed response probabilities for the subset of items relevant for various conditions. This procedure is described in detail in the Appendix; we summarize it briefly here. Participants in Experiments 1 and 2 were limited to four possible fixation targets on any trial. Although the display limits the responses we can observe, we do not assume that lexical activation was limited to the displayed alternatives (given, for example, the results shown in Figure 5 and evidence with real words that nondisplayed competitors influence the time course of fixation [Dahan, Magnuson, Tanenhaus, & Hogan, 2001]). To explicitly link the model's output with the task faced by participants, we assume lexical activation depends on all the items in the artificial lexicon, but a choice must be made from the displayed alternatives. Thus, response probabilities for the four possible alternatives are computed using a variant of the R. D. Luce (1959) choice rule.

Figure 8 plots the changes in the predicted response probabilities to targets, cohorts, rhymes, and unrelated distractors over training (see the Appendix for details of how activations were converted to response probabilities). The top left panel of Figure 8 shows response probabilities very early in training, after 540 epochs. The next five panels show cohort and rhyme effects as training progresses.[10]

Note that the response probability for the rhyme is comparable with or greater than that of the cohort item in the first two panels. In the second panel (720 epochs), we observe perhaps the closest fit to the data from Experiments 1 and 2, in which the rhyme is active relatively early and reaches a maximum similar to that of the cohort. In the third and fourth panels, we see patterns similar to those of Allopenna et al. (1998; both in their human eye movement data and their simulations with TRACE). In the fourth and fifth panels, we see the network's behavior as error rate is reaching asymptote, and rhyme effects eventually disappear. Thus, the present simulations provide a suggestive explanation for the change in relative magnitudes of cohort and rhyme effects in Experiments 1 and 2: Even at Epoch 540, there is an initial advantage for the cohort over the rhyme (e.g., from Cycles 5 to 10). This is consistent with the second explanation for the shift in competitor effects: Processing is incremental, and the advantage for overall similarity compared with onset similarity early in

[10] There is one significant characteristic of all the simulations shown that is not observed in our human data: Target response probabilities drop sharply during the last few time slices of the final vowel. This is because the model is anticipating the next word. By reducing the activation of the current word, error can be minimized after its offset (i.e., at the onset of the next word).

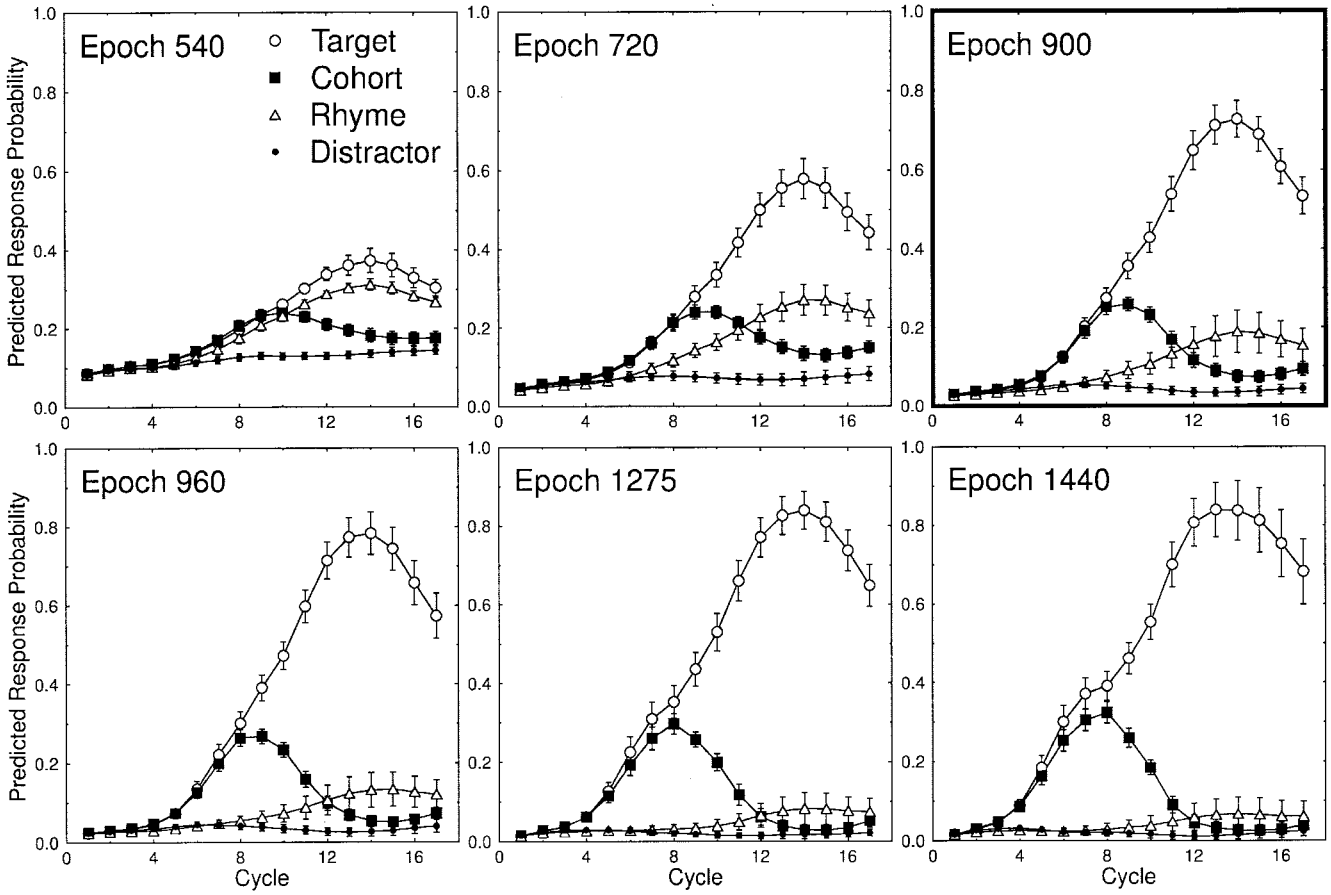

Figure 8. The developmental time course of competitor effects in the simulations. Initially, rhymes compete more strongly than cohorts. As the network learns the statistics of the training corpus, rhyme effects disappear. Epoch 900 is highlighted because it is the epoch selected for further analysis (see Figure 9). The error bars show 95% confidence intervals.

training reflects weak representations rather than holistic representations.

Figure 9 shows that midway through training (after 900 epochs), the network captures the major trends we found in Experiments 1 and 2. The top left panel shows that the basic pattern (averaging over frequency conditions) holds. The bottom left panel shows that similar results are found when targets and competitors are the same frequency. The bottom right panel shows that the advantage found for HF competitors over LF targets is predicted by the network. Finally, the top right panel provides a comparison to the absent competitor condition from Experiment 2. The large circles represent the predicted response probability for HF targets that either have LF (filled circles) or HF (open circles) competitors, when those competitors are absent from the decision process. To accomplish this, we took the activations of four items into account in the decision rule described in the Appendix: the target and three randomly selected distractors (aside from the target's cohort and rhyme). Even when the competitors are not included in the decision rule (by analogy to their not being displayed in Experiment 2), they have strong effects on targets; the predicted response probability of targets with HF competitors rises more slowly than that of targets with LF competitors.

Discussion

The simulations capture the major posttraining trends observed in Experiments 1 and 2, including cohort and rhyme competition, and the modulation of those effects by target and competitor frequency (i.e., neighborhood density effects). Although the fit between the model and data is not perfect (e.g., the lag between cohort and rhyme effects is too long in the model, and target probabilities drop near word offset), the simulations show that SRNs are compatible with the basic phenomena of spoken word processing. The discrepancies between the model and data likely depend largely on the nature of the input representations used. Rather than tweaking the current representational scheme to fit the data more closely, in future work we will use more realistic input (e.g., along the lines of the representations used by Plaut & Kello, 1999).

The SRN also provides a potential explanation for the change in cohort and rhyme effect magnitudes between Days 1 and 2 of training. As discussed above, the model is consistent with an explanation of the shift that depends on incremental processing throughout learning, rather than a shift from holistic to incremental

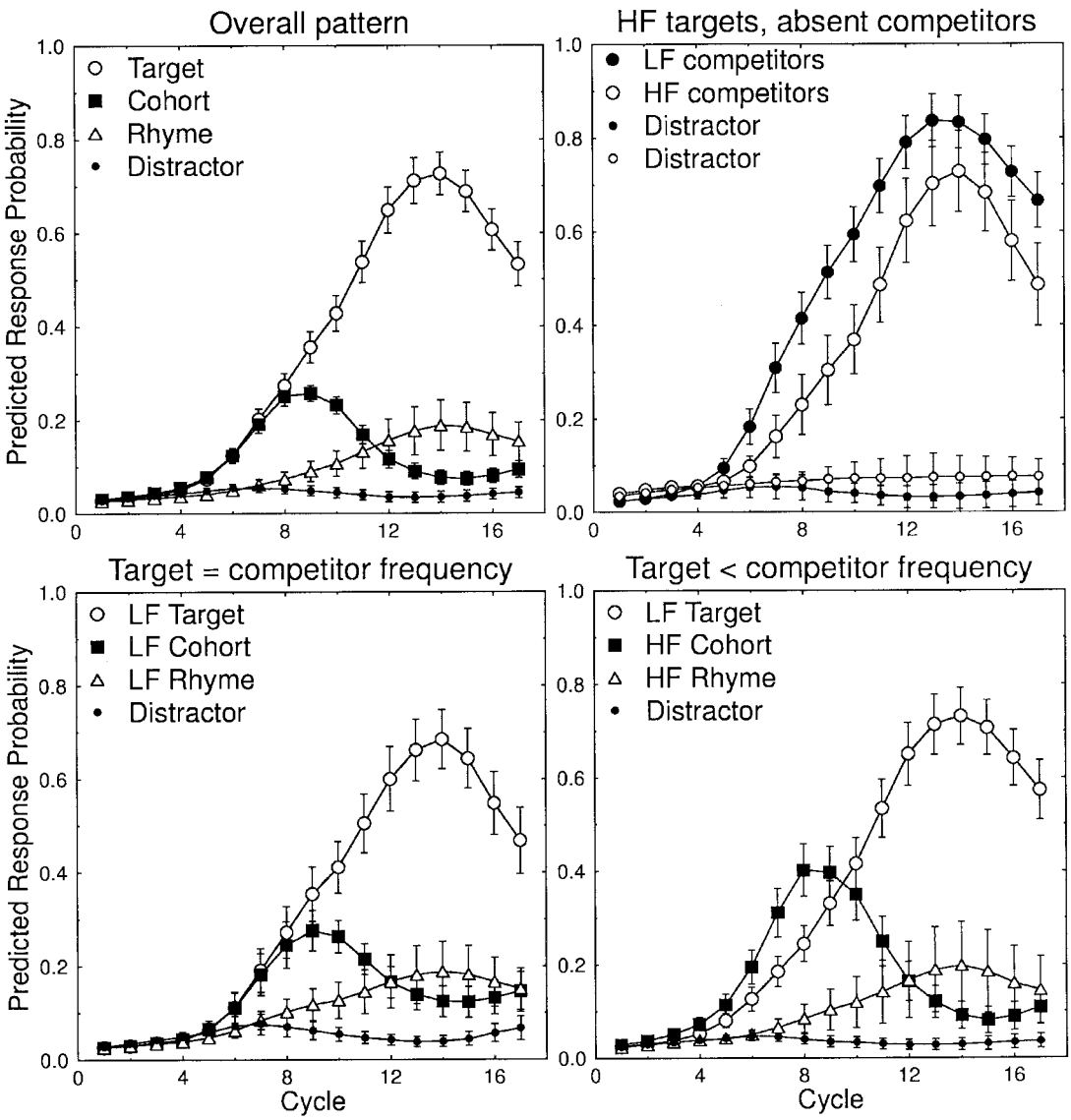

Figure 9. Simulations of the major trends from Experiments 1 and 2 after 900 training epochs. Top left: overall pattern (averaged over all frequency conditions). Top right: effects of competitors even when they are "absent," that is, not included in the decision rule. Bottom left: competitor effects given equivalent target and competitor frequencies (all low frequency [LF], in this example). Bottom right: competitor effects given an LF target and high-frequency (HF) competitors. The error bars show 95% confidence intervals.

processing.[11] This has implications for the hypothetical holistic-to-incremental shift hypothesis in the developmental literature (e.g., Charles-Luce & Luce, 1990; Ferguson & Farwell, 1975), suggesting that the basis for apparently holistic processing may be weakly learned lexical representations.[12] Indeed, this explanation makes a very specific prediction: The items children seem to process holistically should be recently learned or infrequent words. We return to this issue in the General Discussion.

Experiment 3

Experiments 1 and 2 demonstrated the feasibility of using artificial lexicons to test specific hypotheses with precisely controlled stimuli. The stimuli in those experiments were designed to have minimal semantic relatedness to any real object (by making them refer to novel objects), precisely controlled phonological neighborhoods within the artificial lexicon, and minimal phonological similarity to real words. We did not explicitly test, however,

[11] We cannot conclude that our data could not be modeled by using an explicit holistic-to-incremental shift. However, the slight, early advantage for cohorts seen in the top panel of Figure 2 suggests that a model that is incremental even early in learning may provide the more parsimonious account. We leave this question for future research.

[12] One caveat must be made here, however: A large difference between our simulations and the task faced by participants was that the network learned only the artificial lexicon. Preliminary simulations suggest that the

whether the artificial lexical items were being processed as though they had been added to the English lexicon or if they were functionally isolated from it. While it might be feasible to continue creating artificial lexicons using the heuristic that artificial words should be dissimilar from English words, Experiment 3 tests directly whether native-language lexical items intrude on the processing of artificial lexical items. If an artificial lexicon can be considered self-contained, design constraints would be tremendously reduced. If the native-language lexicon does affect performance on items in an artificial lexicon, one must take great care in designing artificial lexicons to ensure that observed effects are not due to interactions with items in a participant's native-language lexicon.

There are at least two possible bases for the hypothesis that there might be interactions with the native-language lexicon. First, given that the artificial lexicon is being presented in an English carrier phrase (e.g., "click on the pibu"), we might expect that the novel words are simply being added to the native lexicon. Second, even if they are somehow being represented distinctly from English words, Spivey and Marian (1999) found that bilingual participants experienced competition between cross-linguistic cohorts (e.g., Russian "marku" and English "marker") in experiments conducted in only one of the languages.

However, there are also at least two possible bases for the opposite hypothesis. The artificial lexicon might be functionally self-contained because it is a *closed set*. For example, an initial disadvantage for LF items dissipates when items are repeated in an experiment (e.g., Pollack, Rubenstein, & Decker, 1959), suggesting that when participants (implicitly or explicitly) develop strong expectations about the possible stimuli, they are able to constrain consideration to the expected set. A second possible basis is *recency*. The many recent presentations of the artificial items may boost their saliency (potentially through enhanced resting level activation, for example) such that the representations of native-language lexical items compete negligibly with the artificial items. Support for this view comes from recent work examining the learning of novel phonotactic constraints in production (Dell et al., 2000) and in comprehension (Chambers et al., 2003; Onishi et al., 2002). Dell et al. found that after brief experience reading aloud lists of CVC syllables with specific phonotactic consistencies, participants' speech errors in a subsequent task conformed to those consistencies. Onishi et al. found that participants who heard similar lists of items showed facilitation on naming items that conformed to the consistencies in the exposure phase, although the effect was short-lived—the advantage quickly dissipated as participants were exposed to items that violated the exposure phase consistencies. Chambers et al. found similar results with infants.

If we do not find effects of English neighborhood density on artificial lexical items, we will not be able to distinguish between closed set and recency explanations. Our present purpose, however, is simply to determine whether it is likely that effects observed with artificial lexicons could be due to characteristics of the native-language lexicon.

In Experiment 3, we examined the potential influence of the native lexicon on a learned artificial lexicon by creating novel words that, if they were English words, would vary in their neighborhood density. Half would fall into high-density neighborhoods, and half would fall into low-density neighborhoods. Half of the items that would be in high-density neighborhoods and half that would be in low-density neighborhoods were presented with HF in the artificial lexicon training, and half were presented with LF. If the newly learned lexicon is self-contained, we should only observe effects of the artificial lexicon's structure (i.e., a frequency effect). If the native-language lexicon influences recognition of the newly learned lexicon, we should observe an interaction of artificial and English lexical effects. For example, if the artificial lexical items are competing for recognition with English lexical items, LF artificial words that would be in high-density English neighborhoods should be harder to recognize than LF artificial words that would be in low-density English neighborhoods.

Method

Participants. Nine native speakers of English who reported normal or corrected-to-normal vision and normal hearing were paid for their participation. Participants attended sessions on 2 consecutive days and were paid $7.50/hr.

Materials. The linguistic materials consisted of 20 artificial words formed by selecting LF low-cohort density (where cohort density was defined as the summed log frequency of items beginning with the same initial CV) and high- and low-neighborhood density English words, and changing the final consonant. Half of the resulting artificial words fall into high-density English neighborhoods, whereas the other half fall into low-density English neighborhoods (see Table 9). The low- and high-density real-word cohorts listed in Table 9 are subsets of items used in a study comparing effects of frequency, neighborhood, and cohort density with real words (Magnuson, 2001, Experiment 4). In that study, LF low-cohort items from low-density neighborhoods had significantly higher fixation proportions than LF low-cohort items from high-density neighborhoods, $F(1, 21) = 18.9$, $p < .001$, $\omega^2 = .29$. This was a strong effect, and power was .98. We estimated that with 9 participants power would be .72, and therefore, if English neighborhood density exerted a similar influence in the artificial lexicon paradigm, we would be highly likely to detect such an effect.

The auditory stimuli were produced by a male native speaker of English in a sentence context ("Click on the yarp."). The stimuli were recorded using a Kay Lab CSL 4000 with 16-bit resolution and a sampling rate of 22.025 kHz. We measured the mean durations of the "Click on the . . ." and the target (artificial) word in each stimulus in the low- and high-English density sets and submitted the measures to *t* tests to ensure there were no reliable differences that might influence processing. The differences were not reliable for the "Click on the . . . " portion of the instructions (low density [$M = 408$ ms, $SD = 20$] vs. high density [$M = 404$, $SD = 17$]), $t(18) = 0.5$, $p = .6$, or for the words (low density [$M = 517$ ms, $SD = 76$] vs. high density [$M = 494$, $SD = 40$]), $t(18) = 0.9$, $p = .4$.

The visual materials consisted of 20 unfamiliar shapes constructed by randomly filling 18 contiguous cells of a 6×6 grid. A distinctive set was generated by creating 500 such figures and randomly selecting 20. Nine examples are shown in Figure 10. Pilot tests indicated that these materials, while clearly similar to those used in Experiments 1 and 2, were more distinctive and easier to learn, presumably because of their increased complexity and thus lower similarity.

Procedure. Participants were trained and tested in sessions on 2 consecutive days. Each session lasted between 90 and 120 min. On Day 1, participants were trained on a 2AFC task for four blocks, and then on a 4AFC task for seven blocks. On Day 2, training continued with seven 4AFC blocks. At the end of each day, participants were given a 4AFC test with no feedback. Eye movements were tracked during the test. One change in the methodology from Experiments 1 and 2 was that we did not instruct participants to fixate a central cross at the beginning of the trial.

---

same apparent holistic-to-incremental shift is observed with an SRN already trained on another set of words. These preliminary simulations have used only a small number of items, however, and we leave this question for future research as well.

Table 9
Stimuli Used in Experiment 3

Item	Gloss	Phonemic	English cohort	No. of NBs	NB density	No. of cohorts	Cohort density
			Low density				
LD 1	fahv	fav	fox	9	24.80	57	87.41
LD 2	goodge	ɡuj	goose	7	9.67	8	6.38
LD 3	hoon	hʊn	hook	10	16.13	9	11.45
LD 4	kef	kɛf	keg	11	20.08	29	44.22
LD 5	kowg	kaʷɡ	couch	4	8.19	35	61.28
LD 6	sheb	ʃɛb	chef	10	21.79	17	25.24
LD 7	thuz	θʌz	thumb	8	11.31	10	12.99
LD 8	torl	tcrl	torch	2	3.00	27	35.68
LD 9	vysh	vaⁱʃ	vice	4	5.32	24	45.42
LD 10	yarp	yarp	yarn	7	13.66	11	15.50
LD means				7.2	13.39	22.9	34.56
			High density				
HD 1	buut	bʊt	bull	35	94.48	48	59.47
HD 2	chihs	çɪs	chick	28	57.84	28	39.23
HD 3	goen	ɡon	goat	36	88.59	27	38.47
HD 4	kayd	ked	cake	40	78.69	38	61.65
HD 5	nide	naⁱd	knight	36	92.40	37	51.16
HD 6	naik	nek	nail	37	91.45	24	50.34
HD 7	nuch	nʌç	nun	22	61.68	22	35.62
HD 8	sahn	san	sock	46	109.42	52	72.24
HD 9	sheed	ʃid	sheep	38	89.56	14	26.69
HD 10	vait	vet	vase	31	88.87	13	24.43
HD means				34.9	85.30	30.30	45.93

Note. NB = neighborhood; LD = low density; HD = high density.

We have adopted this approach in recent research (e.g., Dahan, Magnuson, & Tanenhaus, 2001; Dahan, Magnuson, Tanenhaus, & Hogan, 2001) to decrease reminders to participants that we are monitoring their eye movements. The only difference we have observed in the nature of eye movement data collected under this method is that the probability of fixating any one of the four displayed items at the onset of the target name approaches .25 instead of 0.

The structure of the training trials was nearly identical to that used in Experiments 1 and 2. First, a central fixation square appeared on the screen. The participant then used the computer mouse to click on the fixation square to begin the trial. After 100 ms, either two shapes (in the first four training blocks) or four shapes (in the rest of the training blocks and the tests) appeared (see Figure 1 for analogous examples of displays from Experiment 1). In contrast to Experiments 1 and 2, participants were not given explicit instructions to fixate the central stimulus. When the participant clicked on the fixation square, a 100-ms pause was followed by the appearance of the pictures and the spoken instruction (e.g., "Click on the yarp."). When participants responded, all of the distractor shapes disappeared, leaving only the correct referent. The name of the shape was then repeated. The object disappeared 200 ms later, and the participant clicked on the square to begin the next trial. The testing block was identical to the four-item training, except that no feedback was given (150 ms after the participant clicked on an object, all of the pictures disappeared).

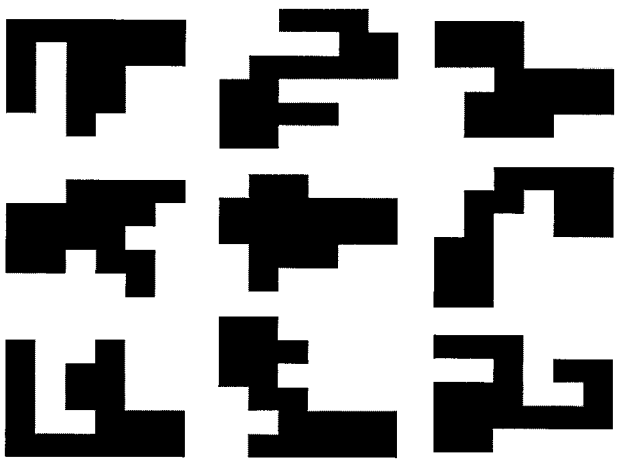

Figure 10. Examples of visual stimuli used in Experiment 3.

During training, half of the items were presented with HF, and half with LF. Frequency assignments were made randomly for each participant, with the constraint that the items in each frequency level were composed of equal numbers of items from each English neighborhood density level. HF items were presented six times per training block, and LF items were presented once per block, so there were 70 trials per training block. Each item was presented six times in each test. For training and testing, distractors were chosen randomly. Trials were presented in random order, with the constraint that the same target could not occur on consecutive trials.

During the tests, eye movements were monitored using a SensorMotorics 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 automatically from point-of-gaze data using SMI's software. The automated saccade detection software used the following parameters: saccade velocity threshold of 30°/s, saccade acceleration threshold of 8,000°/s^2, and saccade motion threshold of .1° (i.e., apparent saccades smaller than .1° were ignored). Mean calibration error across both days and all participants was .41° (*SD* = .10, range = .20 to

.60), which was considerably smaller than the size of the pictures (approximately 1°) and the distance between pictures (approximately 2°). The auditory stimuli were presented binaurally through headphones (Sennheiser HD-570) using standard Macintosh Power PC digital-to-analog devices.

Predictions

First, we expect to observe an effect of training frequency. HF words during training should be processed more readily than LF words, which should be reflected in a more rapid rise in fixation proportions for HF words. Second, if there is intrusion from the English lexicon—that is, if English words compete for recognition with the artificial lexical items—words that would fall into high-density English neighborhoods should be harder to recognize than items that would fall into low-density neighborhoods. The strong version of this interaction prediction is that we ought to find a reliable density effect or a reliable Density × Target Frequency interaction. A weaker version is that we ought to find stronger effects of target frequency on items that would fall into low-density neighborhoods or stronger effects of density on items in the artificial lexicon that are LF. We conducted planned comparisons to examine the weaker predictions even if the Density × Target Frequency interaction is not reliable.

Alternatively, if the artificial lexicon is functionally encapsulated from the English lexicon (whether due to recency or membership in a closed set), we should not observe effects of English neighborhood density. Because the novel words selected for the artificial lexicon differed only on the final phoneme from neighbors in English, the present experiment is a particularly stringent test of intrusion from the English lexicon, as competitor activation by any model of spoken word recognition is biased toward word-initial phonemes.

Results

Training. The progression of training accuracy is detailed in Table 10. Participants quickly reached ceiling levels of accuracy on HF items (by about the third 2AFC block), though it took a bit longer to reach ceiling for LF items (about the third 4AFC block). A 4 (block) × 2 (frequency) × 2 (density) ANOVA on Day 1 accuracy in 2AFC blocks revealed significant main effects of block (see selected means in Table 4), $F(3, 24) = 16.3$, $p < .001$, $\omega^2 = .66$, and frequency (HF [$M = .91$, $SD = .11$] vs. LF [$M = .71$, $SD = .25$]), $F(1, 8) = 27.7$, $p < .005$, $\omega^2 = .77$, but not of English density (high density [$M = .82$, $SD = .23$] vs. low density [$M = .80$, $SD = .21$], $F(1, 8) < 1$. Similar trends held for 4AFC blocks on Days 1 and 2, except that the effect of block was not significant for either analysis, because accuracy quickly approached ceiling levels (due to time constraints, 1 participant did not complete all the 4AFC training sessions or the test on Day 1, thus the change in degrees of freedom in the following results). The effect of frequency was reliable in 4AFC blocks on Day 1 (HF [$M = .92$, $SD = .23$] vs. LF [$M = .89$, $SD = .23$]), $F(1, 7) = 21.1$, $p < .005$, $\omega^2 = .74$, but not on Day 2 ($F < 1$), because accuracy was at ceiling levels. The effect of density was not reliable either day ($F < 1$).

Eye-tracking tests. Participants reached ceiling levels of accuracy on their mouse-click responses on both days' tests (accuracy > .99 in all conditions), and there were no significant accuracy effects or interactions. Fixation proportions over time are plotted for the main effects of frequency and density on Days 1 and 2 in Figure 11. From the main effects, it appears that there was an effect of artificial lexical frequency, but only a hint of an effect of English density. However, the effects of density at the low and high levels of frequency are shown in Figure 12. From these results, it appears that the lack of a main effect of density might be masked by an interaction with frequency: At LF, there is an advantage for items that would fall into low-density English neighborhoods, whereas at HF, there is perhaps a slight trend in the opposite direction.[13]

Table 10
Accuracy in Training and Testing in Experiment 3

Block	Overall	HF	LF
Training 1 (2AFC)	.68	.78	.58
Training 4 (2AFC)	.91	.98	.85
Training 5 (4AFC)	.86	.95	.78
Training 11 (4AFC)	.89	.89	.90
Day 1 test	.96	.98	.93
Training 12 (4AFC)	.95	.99	.91
Training 15 (4AFC)	.93	.95	.91
Training 18 (4AFC)	.97	.97	.97
Day 2 test	.98	.98	.99

Note. HF = high frequency; LF = low frequency; 2AFC = two-alternative forced choice; 4AFC = four-alternative forced choice.

We conducted ANOVAs on mean fixation proportions in the window from 200 ms (when we would expect the earliest signal-driven differences in fixation proportions) to 1,400 ms (approximately where target fixation proportions asymptote in each condition). Mean fixation proportion was computed by summing the target fixation proportion at each sample and then dividing by the number of samples. We conducted identical analyses on the data from both days, with separate analyses using participants (F_1) and items (F_2) as the random variable (we report both below; when we report standard deviations, we use those from the analyses by participant, which were always larger than those from the items analyses).

The first analysis was a 2 (day) × 2 (target frequency) × 2 (English density) ANOVA on the proportion of fixations to HF targets presented among MF distractors. The main effect of day was significant, as target fixations increased from .58 ($SD = .11$) to .63 ($SD = .10$) from Day 1 to Day 2, $F_1(1, 7) = 7.4$, $p = .029$, $\omega^2 = .29$; $F_2(1, 36) = 8.5$, $p = .006$, $\omega^2 = .09$. There was also a main effect of target frequency (HF = .63, $SD = .08$; LF = .57, $SD = .12$), $F_1(1, 7) = 8.8$, $p = .021$, $\omega^2 = .33$; $F_2(1, 36) = 12.6$, $p = .001$, $\omega^2 = .13$. The main effect of density was not reliable (high density = .60, $SD = .11$; low density = .61, $SD = .11$), $F_1(1, 7) = 0.39$, $p = .552$, $\omega^2 = 0$; $F_2(1, 36) = 1.4$, $p = .243$, $\omega^2 = 0$, and neither were any of the interactions (Fs < 2.6, ps > .12).

[13] There were differences in the frequency effect at low and high density (not shown), with a stronger frequency effect at high density and only a very weak effect at low density. However, we will not discuss these trends here, because they are of little interest; they are weak and in the opposite direction of what one would predict (that the frequency effect should be stronger at *low* density).

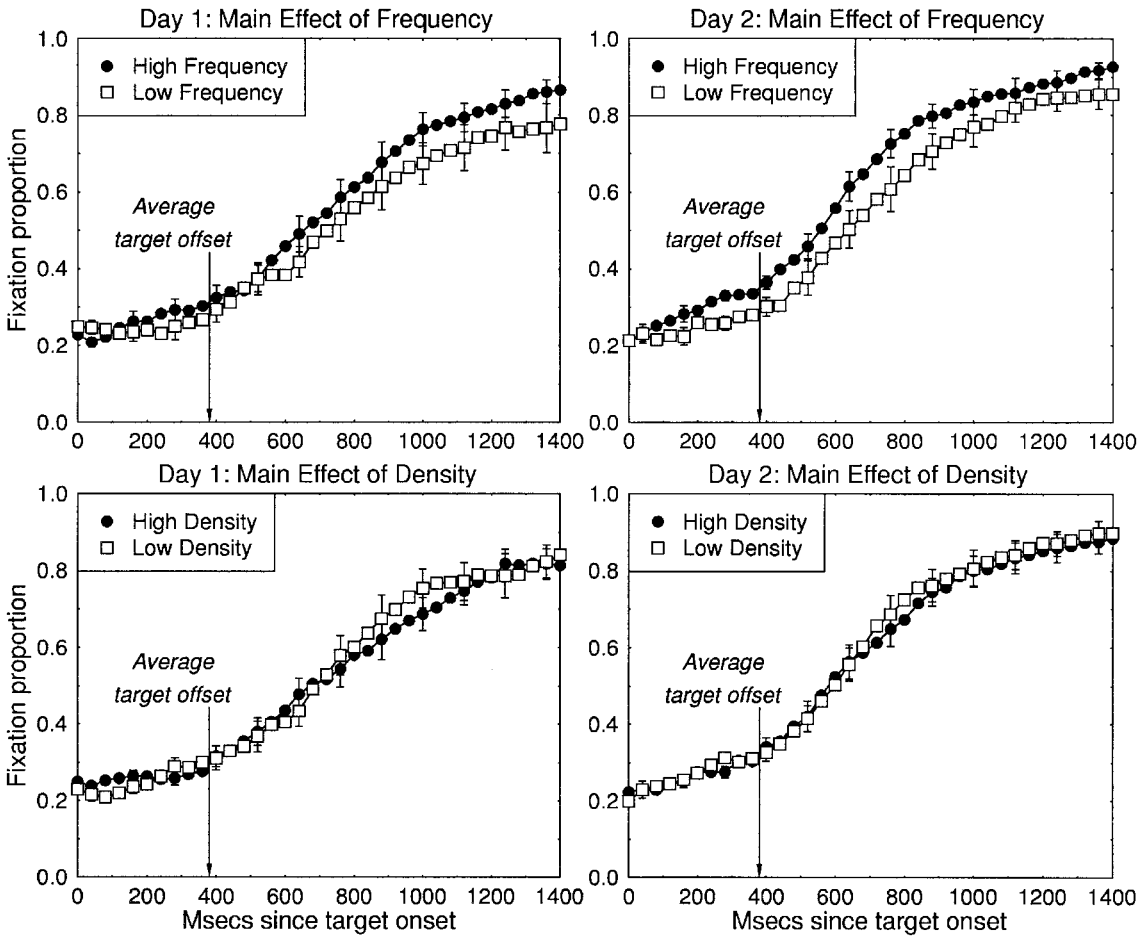

Figure 11. Main effects of frequency (top) and density (bottom) on Days 1 (left) and 2 (right) in Experiment 3. Standard error bars are shown for every third data point.

Thus, the results of the overall analysis suggest that there is not an interaction between artificial lexicon frequency and native-language lexical density.

Recall, however, the weaker predictions for the interaction hypothesis: differential effects of target frequency at high and low density (specifically, a stronger effect at low density) and differential effects of density at HF and LF (a stronger effect at LF). Analyses of simple effects revealed differential effects, but the pattern was not consistent with the predictions. The effect of density was not reliable at HF (high density = .64, SD = .08; low density = .62, SD = .08), $F_1(1, 7) = 1.8$, $p = .22$, $\omega^2 = .04$; $F_2(1, 36) = 0.3$, $p = .614$, $\omega^2 = 0$. The effect was not reliable at LF in the participants analysis (high density = .55, SD = .11; low density = .59, SD = .13), $F_1(1, 7) = 1.8$, $p = .23$, $\omega^2 = .05$, but was marginally reliable in the items analysis, $F_2(1, 36) = 0.37$, $p = .064$, $\omega^2 = .03$. Note that the trend was in the predicted direction for LF but not for HF. This is partly consistent with a weak interaction prediction, but not convincing given the weakness of the effect. The effect of target frequency was reliable at high density (HF = .64, SD = .08; LF = .55, SD = .11), $F_1(1, 7) = 9.8$, $p = .017$, $\omega^2 = .35$; $F_2(1, 36) = 13.3$, $p = .002$, $\omega^2 = .13$, but not at low density (HF = .62, SD = .08; LF = .59, SD = .13), $F_1(1, 7) = 0.48$, $p = .511$, $\omega^2 = 0$; $F_2(1, 36) = .70$, $p = .414$, $\omega^2 = 0$. Note that this latter pattern—a stronger target frequency effect at high than low density—is inconsistent with the weak interaction prediction of a stronger effect at low than high density. Thus, the overall analysis and the planned comparisons of simple effects lead to the same conclusion: There is scant evidence for even weak interaction between artificial lexicon frequency and English neighborhood density.

The planned separate analyses of each day revealed similar patterns. On Day 1, only the effect of target frequency approached significance (HF = .60, SD = .09; LF = .55, SD = .13), $F_1(1, 7) = 3.6$, $p = .10$, $\omega^2 = .14$; $F_2(1, 18) = 2.2$, $p = .152$, $\omega^2 = .03$ (all other effects and interactions, Fs < 1.8, ps $> .22$, ω^2s $< .05$). Of the simple effects of target frequency and density, the only reliable effect was target frequency at high density, as in the overall analysis. On Day 2, the only reliable effect was that of target frequency (HF = .62, SD = .13; LF = .56, SD = .14), $F_1(1, 8) = 7.0$, $p = .03$, $\omega^2 = .25$; $F_2(1, 18) = 6.1$, $p = .024$, $\omega^2 = .11$ (all other effects and interactions, Fs < 2.1, ps $> .19$, ω^2s $< .056$). Again, the only reliable effect among the simple effects of target frequency and density was target frequency at high density.

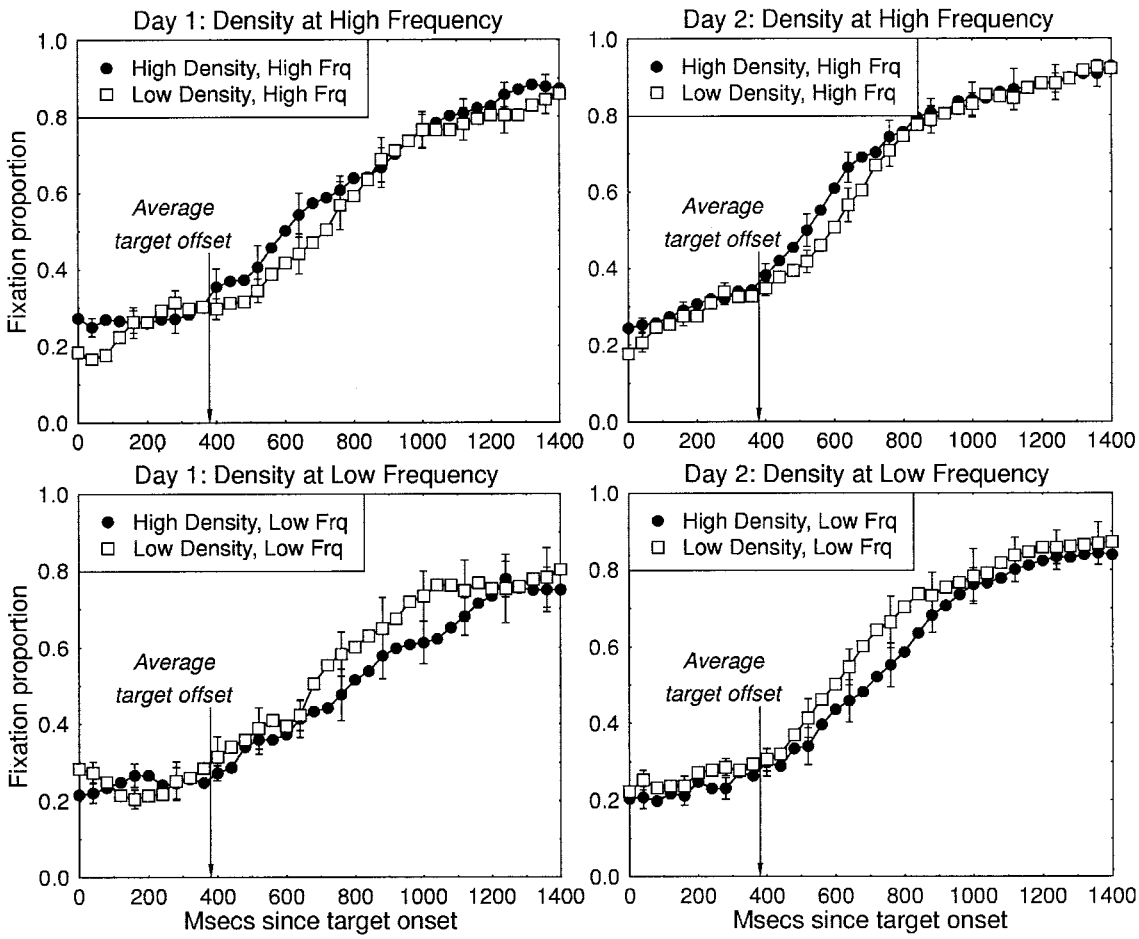

Figure 12. Neighborhood density effects at high (top) and low (bottom) levels of frequency (Frq) on Days 1 (left) and 2 (right) in Experiment 3. Standard error bars are shown for every third data point.

Some caution is in order before we accept the null hypothesis that density does not affect responses to words in the artificial lexicon. Two aspects of the data suggest that the effect of density at LF requires further analysis. First, it is apparent from Figure 12 that the influence of density occurs in a smaller window than the one we have used in analyzing all three experiments. The time window used to analyze the study with comparable real words on which we based the power estimate in the *Materials* section (Magnuson, 2001, Experiment 4) was from 200 ms after word onset through 1,000 ms, because fixation probabilities typically asymptote by 1,000 ms with real words. However, this suggests we might only expect to detect English density effects in the smaller time window. Second, there was a relatively large effect size in the participants analysis of density at LF. Given that we are looking for interactions with real words, we might expect their influence to conform to the time scale observed with competition between real words, and so we repeated our analyses on the window from 200 to 1,000 ms. These analyses closely resemble those with the longer time window, so we focus on points of divergence.

In the overall analysis, the only difference was a strong trend toward a Target Frequency × Density interaction, $F_1(1, 7) = 4.2$, $p = .079$, $\omega^2 = .17$; $F_2(1, 36) = 3.7$, $p = .062$, $\omega^2 = .03$. In the analyses of simple effects, target frequency was reliable only at high density (HF = .54, *SD* = .08; LF = .44, *SD* = .10), $F_1(1, 7) = 11.2$, $p = .012$, $\omega^2 = .39$; $F_2(1, 36) = 13.3$, $p = .002$, $\omega^2 = .13$. Again, this is inconsistent with the interaction hypothesis, which predicts a stronger effect at low density, not high density. However, there was an important change in the effect of density: It was still not reliable at HF (high density = .54, *SD* = .08; low density = .50, *SD* = .08), $F_1(1, 7) = 2.9$, $p = .13$, $\omega^2 = .10$; $F_2(1, 36) = .52$, $p = .477$, $\omega^2 = .0$, but there was a strong trend in the predicted direction at LF (high density = .44, *SD* = .10, low density = .49, *SD* = 12), which was marginally reliable in the participants analysis, $F_1(1, 7) = 3.8$, $p = .094$, $\omega^2 = .15$, and reliable in the items analysis, $F_2(1, 36) = 4.7$, $p = .036$, $\omega^2 = .04$. This pattern is consistent with the weak interaction hypothesis, because there was an effect at LF but not HF.

In summary, we found reliable effects of artificial lexical frequency throughout the time course of target fixation. Although there was not a strong trend toward a main effect of density, we did find differential effects of density at LF and HF. When we constrained the window of analysis to the 200–1,000 ms used in a comparable study with real words (Magnuson, 2001, Experiment 4), we found that mean fixation proportion was reliably higher for

LF artificial lexical items that would fall into low-density English neighborhoods than for items that would fall into high-density neighborhoods.

Discussion

The main effects from the eye-tracking test suggest that, under conditions like those in Experiments 1–3, an artificial lexicon shows little intrusion from a participant's native lexicon, especially when compared with the magnitude of the frequency effect that resulted from artificial lexical training. There was a significant effect of the experimental frequency manipulation, but not of English neighborhood density. While we cannot distinguish between the two possible bases discussed earlier for this pattern (closed set vs. recency), the purpose of Experiment 3 was simpler. We wished to test whether characteristics of the native-language lexicon impinge on an artificial lexicon like those used in Experiments 1 and 2. A word of caution is warranted, however, because there was a trend toward an effect of English density on LF words that became significant when we shortened the window of analysis. It is possible that LF items in an artificial lexicon may be susceptible to native lexicon intrusion effects, especially early in training. Indeed, the effect appeared weaker after the 2nd day of training. All the same, the materials for Experiment 3 were designed to be extremely similar to English words, and it is notable that (a) we did not find a reliable main effect of density comparable with that found with similar nonwords (Magnuson, 2001), and (b) we did not find reliable effects of English density on HF artificial lexical items.[14]

This suggests that experimenters can expect negligible interference from the native lexicon, even when artificial lexical items closely resemble real words, provided that the artificial items are presented frequently. That is, recent experience dominates, as has been found with subsyllabic regularities (Chambers et al., in press; Dell et al., 2000; Onishi et al., 2002). So an artificial lexicon of words that conform to native phonotactics might be considered functionally isolated from a participant's native lexicon—when the items are presented frequently and there is no exposure to anything but the training materials prior to or during the test phase.

General Discussion

Experiments 1, 2, and 3 and SRN simulations of Experiments 1 and 2 met the four goals laid out in the introduction. We replicated previous competition and frequency effects (Experiments 1 and 2). We found that lexical activation in the eye-tracking paradigm is not constrained to the set of displayed objects (Experiment 2). We were able to account for the major trends in Experiments 1 and 2 using a model (SRN) that also provided an account of the learning aspect of the artificial lexicon task. Finally, Experiment 3 suggests that artificial lexicons may be considered functionally isolated from the native lexicon when the items conform to native phonotactics and have been presented recently and frequently. Even with materials designed to be maximally similar to specific English words, we observed relatively weak effects of English neighborhoods and only LF artificial lexical items. Together, the experiments and simulations demonstrate the viability of the artificial lexicon approach for studying the microstructure of spoken word recognition.

After minimal training, lexical processing in a novel lexicon is highly similar to natural language spoken word recognition. In addition to replicating previous time course effects, we also provided the first on-line time course measures of neighborhood density effects (e.g., the modulation of cohort and rhyme effects by target and competitor frequency). The time course measures showed that after training, the artificial lexical items were processed incrementally. Phonetically similar neighbors became partially active during the processing of spoken words with a time course that mapped onto emerging phonetic similarity. Furthermore, the time course depended on competition among all items in the artificial lexicon; although we explicitly acknowledge the influence of the visual display in the linking hypothesis described in the Appendix, the time course of recognition for items depended on their neighborhood densities, even when their neighbors were not included in the set of displayed pictures.

The experiments also address whether incremental processing is the natural mode of spoken language processing or whether it takes substantial training to support incrementality. After 1 day of training in our task, the magnitudes of cohort and rhyme competition effects were strikingly different from those observed with real words or after a second training session. After the second training session, a strongly incremental pattern was observed. Target and cohort fixation proportions separated together from the unrelated baseline around 200 ms after word onset (as discussed earlier, this approaches the theoretical minimum lag for signal-driven fixations). After an additional lag, usually equivalent to the delay between word onset and the offset of the first syllable nucleus, the rhyme proportion separated from the baseline. Shortly thereafter, the proportion of fixations to the cohort began to fall back to baseline. Thus, fixation proportions closely tracked phonetic similarity over time. After the 1st day of training, however, the cohort and rhyme proportions followed a very similar time course and reached similar peaks. This pattern of data was compatible with two explanations: (a) a holistic to incremental shift and (b) strengthening of lexical representations with learning within a system that is fundamentally incremental.

The SRN simulations were consistent with the second alternative. Early in training, rhyme effects in the network were much larger than cohort effects. All the same, the network was still sensitive to incremental information, as evidenced by an advantage for the cohort during the initial portions of a word. This suggests that the basis for the weak cohort effect early in learning is the relative weakness of lexical representations. As training begins, the network has not learned that initial segments are highly predictive within the lexicon. Thus, local, incremental similarity is more important, with strong additive effects for overall similarity combined with a strong negative impact of mismatching segments at any position. This favors rhymes because of their greater overall similarity. Late in a word, because the network has not yet encoded the relationship between early and late segmental information, rhymes are still considered very likely matches to an input word.

[14] We thank two anonymous reviewers for pointing out that there must be influences of English at work even with our artificial lexical items; for example, violations of English phonotactics would likely have impeded learning. However, the goal of the present experiment was to test for evidence of English lexical competition.

As training progresses and the network learns word-specific sequential dependencies, the advantage shifts to cohorts, for reasons similar to those that underlie the onset advantage in TRACE (though they can be better described in terms of conditional probabilities; the network has implicitly encoded the probability of a series of segments given the preceding series).

This suggests the possibility that children's early lexical representations may not be fundamentally different from adults' but are also incrementally organized. Evidence suggesting holistic processing may instead reflect weak lexical representations. This hypothesis makes strong predictions about the conditions under which apparent holistic processing should be observed, for example, when a word has recently been introduced to the lexicon or is infrequent. This also suggests a way to reconcile reports of holistic processing with several studies that suggest young children and infants are sensitive to fine-grained, incremental phonetic information (e.g., Gerken et al., 1995; Swingley & Aslin, 2000, 2002; Swingley et al., 1999). However, the problems we discussed earlier with estimating factors such as frequency from corpora for adults are even more dramatic for children. Frequency counts in a corpus are difficult to apply when the items predicted to be critical for testing the hypothesis are those that the child is just beginning to learn, no matter their frequency in a corpus of adult speech or text. For the purpose of this hypothesis, frequency must be operationalized as the true frequency with which the child has experienced a particular form. A project under development will examine this issue explicitly by using the artificial lexicon paradigm with children of different ages to precisely control frequency of occurrence. This will put children in a situation comparable with that faced by adults in our study and allow us to make direct comparisons of adult and child word learning.

Conclusions

The present results demonstrate that research with a novel lexicon that builds upon an existing phonological system can be used to evaluate the microstructure of spoken language comprehension. This paradigm offers a valuable complement to more traditional paradigms because it allows for (a) precise experimental control of the distributional properties of the linguistic materials, (b) tests of distribution-based learning hypotheses, and (c) evaluation of processing during early lexical learning. Moreover, the use of artificial lexical items that *refer* to tangible objects, and potential extensions to more complete artificial languages with well-defined semantics, should make it possible to explore the interaction of distributional and referential properties during language processing—issues that would be difficult to address in research with nonreferential artificial languages (because of the difficulty of introducing semantic properties) or with natural language stimuli (because of lack of precise control over distributional properties).

Our results also make several empirical and theoretical contributions. Experiments 1 and 2 provide the first time course measures of neighborhood density effects. Experiment 2 demonstrates that lexical activation in the eye-tracking paradigm is not constrained to the displayed items. Experiment 3 suggests that artificial lexicons may be considered functionally isolated from the native lexicon. Simulations with SRNs captured the major post-training trends, as well as the changes in the relative magnitude of cohort and rhyme competition as words were learned. This is consistent with the hypothesis that children's and adults' lexical representations do not differ fundamentally. Evidence consistent with holistic lexical representations is also consistent with a change in the strength of representations as words are learned rather than a change in the nature of the representations. The latter hypothesis leads to specific predictions regarding when processing may appear relatively holistic (e.g., early in learning, especially for infrequent words). Taken together, the present experiments and simulations establish a methodology for examining the time course of the processing of spoken words that enables precise control over stimulus and lexicon properties and provides time course data for testing fine-grained predictions of spoken word recognition models.

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.

Aslin, R. N., Woodward, J., LaMendola, N., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 117–134). Mahwah, NJ: Erlbaum.

Braine, M. D. S. (1963). On learning the grammatical order of words. *Psychological Review, 70,* 323–348.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. P. (1995). Bottom-up connectionist modelling of speech. In J. P. Levy, D. Bairaktaris, J. A. Bullinaria, & P. Cairns (Eds.), *Connectionist models of memory and language* (pp. 289–310). London: University College London Press.

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition, 87,* B69–B77.

Charles-Luce, J., & Luce, P. A. (1990). Similarity neighborhoods of words in young children's lexicons. *Journal of Child Language, 17,* 205–215.

Charles-Luce, J., & Luce, P. A. (1995). An examination of similarity neighbourhoods in young children's receptive vocabularies. *Journal of Child Language, 22,* 727–735.

Coady, J. A., & Aslin, R. N. (in press). Phonological neighborhoods in the developing lexicon. *Journal of Child Language.*

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.

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

Cole, R. A., & Jakimik, J. (1980). A model of speech perception. In R. A. Cole (Ed.), *Perception and production of fluent speech* (pp. 133–163). Hillsdale, NJ: Erlbaum.

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

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

Cooper, R. M. (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.

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

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes, 16,* 507–534.

Dell, G. S., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science, 17,* 149–195.

Dell, G. S., Reed, K. D., Adams, D. R., & Meyer, A. S. (2000). Speech errors, phonotactic constraints, and implicit learning: A study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1355–1367.

Dollaghan, C. A. (1994). Children's phonological neighborhoods: Half empty or half full? *Journal of Child Language, 21,* 257–271.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14,* 179–211.

Ferguson, C. A., & Farwell, C. B. (1975). Words and sounds in early language acquisition. *Language, 51,* 419–439.

Fischer, B. (1992). Saccadic reaction time: Implications for reading, dyslexia and visual cognition. In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and reading* (pp. 31–45). New York: Springer-Verlag.

Fougeron, C. A., & Keating, P. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America, 101,* 3728–3740.

Fowler, C. A., & Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language, 26,* 489–504.

Francis, W. N., & Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar.* Boston: Houghton-Mifflin.

Garnica, O. K. (1973). The development of phonemic speech perception. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 214–222). Hillsdale, NJ: Erlbaum.

Gerken, L., Murphy, W. D., & Aslin, R. N. (1995). Three-and four-year-olds' perceptual confusions for spoken words. *Perception & Psychophysics, 57,* 475–486.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1166–1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105,* 251–279.

Gomez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences, 4,* 178–186.

Howes, D. H. (1957). On the relation between the intelligibility and frequency of occurrence of English words. *Journal of the Acoustical Society of America, 29,* 296–305.

Jordan, M. I. (1986). *Serial order: A parallel distributed processing approach* (Institute for Cognitive Science Report 8604). San Diego: University of California, San Diego.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Lehiste, I. (1970). *Suprasegmentals.* Cambridge, MA: MIT Press.

Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception & Psychophysics, 62,* 615–625.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19,* 1–36.

Luce, P. A., Pisoni, D. B., & Goldinger, S. D. (1990). Similarity neighborhoods of spoken words. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 122–147). Cambridge, MA: MIT Press.

Luce, R. D. (1959). *Individual choice behavior.* New York: Wiley.

Magnuson, J. S. (2001). *The microstructure of spoken word recognition.* Unpublished doctoral dissertation, University of Rochester, Department of Brain and Cognitive Sciences.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word recognition. *Cognition, 25,* 71–102.

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* (pp. 187–210). Hillsdale, NJ: Erlbaum.

Marslen-Wilson, W., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review, 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.

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

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86,* B33–B42.

McQueen, J. M., Norris, D., & Cutler, A. (1999). Lexical influence in phonetic decision making: Evidence from subcategorical mismatches. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 1363–1389.

Menyuk, P., & Menn, L. (1979). Early strategies for the perception and production of words and sounds. In P. Fletcher & M. Garman (Eds.), *Language acquisition: Studies in first language development* (pp. 49–70). Cambridge, England: Cambridge University Press.

Morgan, J. L., Meier, R. P., & Newport, E. L. (1987). Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Psychology, 19,* 498–550.

Newman, R. S., Sawusch, J. R., & Luce, P. A. (1997). Lexical neighborhood effects in phonetic processing. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 873–889.

Nooteboom, S. G., & Kruyt, J. G. (1987). Accent, focus distribution, and the perceived distribution of given and new information: An experiment. *Journal of the Acoustical Society of America, 82,* 1512–1524.

Norris, D. (1990). A dynamic-net model of human speech recognition. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 87–104). Cambridge: MIT Press.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52,* 189–234.

Onishi, K. H., Chambers, K. E., & Fisher, C. (2002). Learning phonotactic constraints from brief auditory exposure. *Cognition, 83,* B13–B23.

Plaut, D. C., & Kello, C. T. (1999). The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist model. In B. MacWhinney (Ed.), *The emergence of language* (pp. 381–415). Mahwah, NJ: Erlbaum.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. E. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103,* 56–115.

Pollack, I., Rubenstein, H., & Decker, L. (1959). Intelligibility of known and unknown message sets. *Journal of the Acoustical Society of America, 31,* 273–279.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35,* 606–621.

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

Savin, H. B. (1963). Word-frequency effect and errors in the perception of speech. *Journal of the Acoustical Society of America, 35,* 200–206.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed developmental model of word recognition. *Psychological Review, 96,* 523–568.

Shvachkin, N. K. (1973). The development of phonemic speech perception in early childhood (E. Dernbach, Trans.). In C. A. Ferguson & D. I.

Slobin (Eds.), *Studies of child language development* (pp. 91–127). New York: Holt, Rinehart & Winston. (Original work published 1948)

Spivey, M. J., & Marian, V. (1999). Cross talk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science, 10,* 281–284.

Stevens, K. N. (1998). *Acoustic phonetics.* Cambridge, MA: MIT Press.

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition, 76,* 147–166.

Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science, 13,* 480–484.

Swingley, D., Pinto, J. P., & Fernald, A. (1999). Continuous processing in word recognition at 24 months. *Cognition, 71,* 73–108.

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. G. (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research,* 29, 557–580.

Tanenhaus, M. K., & Spivey-Knowlton, M. J. (1996). Eye-tracking [Special issue: A guide to spoken word recognition paradigms]. *Language and Cognitive Processes, 11,* 583–588.

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.

Terken, J., & Hirschberg, J. (1994). Deaccentuation of words representing "given" information: Effects of persistence of grammatical function and surface position. *Language and Speech, 37,* 125–145.

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: Vol. 4. Reviews of oculomotor research* (pp. 253–393). Amsterdam: Elsevier.

Appendix

Simulation Details

Simulation Software

We used Doug Rohde's LENS (light, efficient neural simulator) to conduct the simple recurrent network (SRN) simulations (see http://tedlab.mit.edu/~dr/Lens/) and a variety of in-house programs to analyze the results and implement the linking hypothesis (see below).

Model Parameters

The choices of parameters such as learning rate and number of hidden units were settled on without extensive exploration of the parameter space. However, none of our results depend crucially on the specific values chosen. A learning rate of 0.1 was chosen because it is large enough to allow relatively fast learning without causing the network to continually adjust to temporary overlearning of particular patterns (i.e., changing the weights so much in response to an item that error increases substantially for other items). Similar results were obtained with learning rates as small as .01 or as large as .5. The former setting slows learning considerably, whereas the latter speeds it to a degree that obscures the development of competition effects. A set of 20 hidden units was decided on in similar fashion. With substantially fewer hidden units, learning is slowed, and with only a few hidden units, it is unlikely that the model could learn even the simple corpus we used. In neither case does the value chosen reflect a search for parameters that would best fit our data. Rather, these are values that yielded the general behavior we needed from the model: gradual learning in a reasonable number of epochs. Neither of these choices is crucial for explaining the overall trends.

Noise

Uniform noise in the range –.9, .9 was added to the input units both during training and testing. Adding noise moved our idealized speech inputs a small step closer to real speech conditions, in which the input is characterized by noise and variability, and slowed the time course of learning, as described in the text. Note that noise was not necessary for finding any of the results we report in the Simulations of Experiments 1 and 2 section. Noise did slow learning, making the changes in the relative magnitude of cohort and rhyme effects easier to observe by analyzing a subset of the epochs.

Stimulus Representation

As described in the text, the input at each time slice consisted of a set of binary phonetic features (syllabic, consonantal, sonorant, nasal, anterior, coronal, continuant, delayed release, strident, voicing, aspiration, lateral, high, low, back, round, tense, and reduced). A coarse analog to coarticulation was achieved by ramping the vector corresponding to a phoneme on and off over seven input cycles and beginning each consecutive phoneme one input cycle after the center of the preceding phoneme. The ramping was achieved by multiplying the phonetic feature vector corresponding to a phoneme by the following values over seven time slices: .03, .045, .06, 075, .06, .045, and .03. A wide range of values for the ramp vector could achieve the basic aim, which was to give the network weak predictive information about an upcoming phoneme. With significantly larger values, we found that the coarticulatory information became completely diagnostic, leading to sudden jumps in lexical activation as soon as the coarticulatory information was encountered. With significantly weaker values, learning time increased substantially. However, values within about an order of magnitude yield similar results.

The relative values in the ramp vector (incrementing or decrementing by 50% of the smallest value each time step) do not have a large impact on the results either; a flat vector, without ramping, yields similar results. Ramping was used to make the input more similar to that used in the TRACE model.

We mentioned in the text that vowels were made longer than consonants to reflect durational, amplitude, and salience differences in natural speech (e.g., Stevens, 1998). This was done by repeating vowels but still aligning the onset of the repetition one time step past the center of the first occurrence. For example, given the input item /pibo/, /p/ would be presented from Time Slices 1 to 7; /i/ would be presented from Slices 3 to 9 and from 5 to 11; /b/ would be presented from Slices 7 to 13; and /o/ would be presented from Slice 9 to 15 and from 11 to 17. Note that because of the values used to ramp phonemes on and off, overlapping vowels would not ramp on, off, on again, and then off again. When the vectors were summed, a vowel spanned 9 time slices (a small difference when compared with 7 slices for consonants), with values .03, .045, .09 (.06+.03), .12 (.075 + .045), .12 (.06 + .06), .12 (.045 + .075), .09 (.03 + .06), .045, and .03. Figure 7 illustrates the coding for the input corresponding to /pibo/.

Frequency Levels

As described in the text, we used a 4:3 high-frequency (HF) to low-frequency (LF) ratio. We tested a range of ratios between varying the HF level from 2 to 7 and the LF level from 1 to 6. A small range of combinations (e.g., 2:1, 5:3, 5:4) gave approximately the same behavior we

found with 4:3, but 4:3 gave the best qualitative fit. Because the frequency level best suited for matching the 7:1 ratio used with our experiments with human participants was not of theoretical interest, we did not pursue this issue in detail.

Linking Hypothesis

Lexical output activations were converted to response probabilities using a procedure similar to the one described by Allopenna et al. (1998). Because participants in the experiments could fixate only one of the four displayed objects, response probabilities (i.e., predicted fixation probabilities) were computed at each time step. All words were allowed to be activated and compete for recognition. Response probabilities were computed for the three items of interest (target, cohort, and rhyme) and a randomly selected unrelated baseline item. Response strengths (S) for each item, i, at each time step, t, were computed using Equation 1; a value of 10 was used for the constant, k, which controls separation between items with high and low activations. Response probabilities (R) were computed using Equation 2, which simply normalizes response strengths. An adjusted response probability (P) was computed with a procedure similar to that of Allopenna et al. using Equation 3. This last step kept very low activations from being treated as if they were as predictive as higher activations and roughly captures the lag before initial fixations. See Allopenna et al. for more discussion.

$$S_{it} = e^{ka_{it}} \tag{1}$$

$$R_{it} = \frac{S_{it}}{\sum S_{jt}} \tag{2}$$

$$P_{it} = R_{it}\frac{\max(a_t)}{\max(a)} \tag{3}$$

Note that activations in our SRNs closely resembled the resulting response probabilities. However, in addition to incorporating the task constraints faced by our participants, we needed to compute response probabilities to provide a true analog to the absent competitors condition (by leaving the competitors out of the decision rule). See the text and Figure 9 for details.

Received August 21, 2002
Revision received January 27, 2003
Accepted January 28, 2003 ■