`The Dynamics of Spoken Word Recognition in Second Language Listeners Reveal Native-Like Lexical Processing

Henna A. Shin (hennashi@usc.edu), Brian Bauman (baumanb@usc.edu), Imola X. MacPhee (imacphee@usc.edu), and Jason D. Zevin (zevin@usc.edu)

Department of Linguistics, Department of Psychology, Graduate Neuroscience Program, University of Southern California, Los Angeles, CA 90089 USA

Abstract

Models of spoken word recognition in monolingual, native listeners account for the dynamics of lexical activation of intended words and their phonologically similar "competitors," in terms of continuous, cascaded processing dynamics. Here we explore how the dynamics of spoken word recognition differ for second language listeners. Groups of native Korean speakers (KL1) and native English speakers (EL1) listened to recordings of words in three conditions: phonological overlap at the beginnings of the words (cohort), at the ends of the words (rhyme), or without phonological overlap (unrelated), and used a computer mouse to select the matching stimulus from an array of two pictures. There are many reasons to predict that KL1 participants would differ from EL1 participants; for example, participants with non-native speech sound perception might strategically reduce the contribution of anticipatory processes to avoid committing to an incorrect response and thus demonstrate smaller effects of anticipatory competition (cohort effect). Instead, the results did not reveal any interactions between language background and performance across the cohort, rhyme and unrelated conditions. Nor were effects of similarity related to overall performance on independent tests of speech sound categorization or vocabulary. The results suggest that the cohort and rhyme effects are robust features of proficient second language spoken word recognition, despite demonstrable differences in speech sound recognition.

Keywords: speech perception; lexical processing; word recognition

Introduction

Anticipatory processes have long been understood to play a central role in spoken word recognition. For example, the Cohort model characterized word recognition as a process of serially eliminating lexical candidates based on gradually accumulating evidence (Marslen-Wilson, 1989). Most — if not all — contemporary and later models such as TRACE, MERGE and Shortlist (McClelland & Elman, 1986; Norris, 1994; Norris et al., 2000) incorporate anticipatory processes, despite differing from Cohort (and from one another) in many theoretically important ways.

In these interactive activation models, anticipatory effects emerge naturally from activation dynamics in a network of connected nodes representing different levels of description. For example, lexical nodes are modeled as receiving activation from their constituent phonemes in a cascaded manner — that is, as each phoneme becomes active, this activity is instantaneously passed to the words that contain it. In this way, words consistent with the input can become activated based on early, partial input. When multiple words are equally consistent with the input, this creates competition among candidate words, resulting in slower recognition times.

When a "visual world" is presented in which pictures with overlapping names (such as a "beaker" and a "beetle") are present, eye-movements during response planning (Allopenna et al., 1998) and arm movements during response execution i.e., using a "mouse tracking" paradigm (Spivey et al., 2005), can reveal this competition at work in real time. For example, arm movement trajectories to a "beaker" when a "beetle" is present have a greater arc than trajectories when a "speaker" is present. Trajectories when a rhyme competitor like "speaker" is present, in turn, have a greater arc than trials on which an item with a phonologically unrelated name is present, indicating that phonological similarity influences processing after anticipatory processes could, in principle, have resolved the stimulus ambiguity entirely. These data are well accounted for in the TRACE model, in just the terms we have described so far. Here we ask how the dynamics of spoken word recognition differ for second language (L2) listeners in this same task.

It is well documented that late L2 learners' ability to categorize speech sounds is not native-like, and is best understood as involving transfer of L1 categorization abilities, (Best & Tyler, 2007; Flege, 1999; Kuhl, 2004). Thus, we might imagine that spoken word recognition in L2 would have similar dynamics to what is observed in individuals with relatively poor phonological processing abilities. For example, Desroches et al. (2008) compared dyslexic children to typically developing controls, and found differences in competition with rhyme, but not cohort competitors. In contrast, McMurray et al. (2010) examined individual differences in an adolescent population with substantial variability in phonological processing abilities (including participants with language impairment), and found that poorer scores on language tests were associated with higher levels of late-occurring competition for *both* cohort and rhyme distractors. To the extent that L2 listeners' speech sound categorization is less efficient than typical native listeners', we may expect their lexical processing to resemble these populations.

There are also reasons to suspect that differences in how the lexicon is organized may influence second language word recognition. Many studies have demonstrated interference effects from interlingual homographs (Dijkstra et al 1999) and homophones (Schulpen et al 2003; Lagrou et al. 2011), indicating that words in multiple languages can be activated by the same input. Thus, the dynamics of lexical activation in L2 learners may reflect contributions from a much larger lexicon, with many more potential "neighbors" for every word. This could have the effect of diluting any competition effects, or at least generally slowing down lexical activation by introducing greater competition to all conditions.

Cross-language interference effects have been used to argue that bilinguals cannot selectively inhibit the entire lexicon of the irrelevant language (Kroll & Dussais, 2004). But other aspects of lexical processing may be under strategic control. For example, if second language learners are aware that they experience greater ambiguity in processing speech sounds in their second language, they may strategically reduce the contribution of anticipatory processes to avoid committing to an incorrect response, waiting longer before acting on incoming information. Because the mouse-tracking approach we are applying here measures the dynamics of response execution, smaller effects of cohort competition might be expected to arise either as evidence accumulates for the eventual target of arm movement, or as the a graded decision process is used to continuously guide motor movements (Spivey et al. 2005). Under this scenario, we might expect L2 listeners to produce a smaller cohort effect, but the same or larger rhyme effect compared to native listeners.

Thus, differences in lexical processing arising at multiple levels of description — due to either strategic or constitutive differences in processing dynamics — could plausibly contribute to distinct patterns of performance between second language learners and native listeners in the current task.

Methods

Participants

Twenty-five monolingual native-English speaking (EL1) adults, mean age 22.50 years (SD = 4.46) and 58 native-Korean speaking (KL1) adults with a minimum of five years English language experience, and mean age of 22.90 years (SD = 5.87) participated in this study. For the KL1 participants, the average Age of Arrival (AoA) to the United States was 9.40 years of age (SD = 4.33) and mean Length of Residence (LoR) in the United States was 13.5 years (SD = 6.47).

In order to provide estimates of language proficiency, the participants completed a phonetic discrimination task, a category fluency task, and the synonym, antonym and picture vocabulary subtests of the Woodcock-Johnson Tests of Cognitive Abilities, Third Edition (W-J III). For the phonetic discrimination task, participants completed a categorical AXB task for three English phonetic contrasts, /ba/ - / va/, /da/ - /ða/, and /fu/ - /θu/, and for three series of Korean phonetic contrasts difficult for native English speakers (tense, plain, and aspirated stops), /p*/ - /p/, /t^h/ - /t*/, and /t/ - /t^h/. For the category fluency task, participants were given 45 seconds to list as many items as they could in a given category (e.g. fruits, clothing items). The EL1 participants performed the task in English only and the KL1 participants performed the task in both English and Korean. These measures were collected in the same session as the spoken word recognition experiment. Analysis using two-sample t-test revealed significant differences between groups, with poorer performance in the KL1 group for all but the Korean Phonetic Discrimination (AXB) task (see Table 1).

Table 1: Monolingual and Bilingual Performance Across Language Measures

Task	EL1	KL1	Δ
Vocab	15.92 (0.23)	13.95 (0.27)	1.97**
Fluency	15.36 (0.69)	13.03 (0.40)	2.33**
English AXB^	0.97 (0.01)	0.92 (0.01)	0.04*
Korean AXB^	0.88 (0.01)	0.94 (0.01)	-0.06**

Scores are presented as means with standard error in parentheses; Vocab: Mean of all three vocabulary measures; ^Mean of all three contrasts; * $p < 0.05$, ** $p < 0.01$.

Stimuli

Stimuli and other major aspects of the experimental design were taken directly from Allopenna et al. (1998). Each stimulus word was two syllables. Stimuli had a mean duration of 0.578 seconds. The words were recorded in isolation and presented without any carrier phrase using Praat software (Version 5.4.04). Each stimulus was saved in a monaural 44,100 Hz WAV file with 10 ms silence before and after the word. Images for the twenty-four visual stimuli, adapted from Allopenna et al. (1998), were prepared from public domain images found on the internet and resized to 275 x 300 (pixels) in Microsoft Paint.

Procedure

The experiment was controlled using Paradigm (Perception Research Systems, Lawrence, KS) on the Windows 7 operating system. The display resolution was set to 1024 x 768. Mouse movements were recorded at a sampling rate of 125 Hz.

The visual display consisted of one target and one distractor, as seen in Figure 1. Appearance of the target on the top left or top right corner was counterbalanced so the target appeared on each side equally often. Presentation of visual stimuli was randomized with a 500 ms delay between trials.

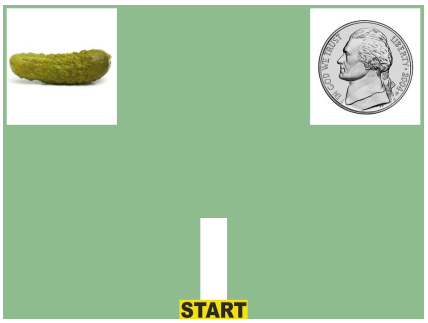

Figure 1. Visual stimulus display during the spoken word recognition task for "pickle" and "nickel".

Familiarization

Participants began with a familiarization phase in which they learned the correspondences between the specific pictures used in the experiment and their (written) names, and practiced moving the mouse continuously throughout the trial as instructed by the experimenter. Familiarization occurred with visual stimuli only, so as not to interfere with the speech perception focus of the task. The goal was to neutralize any potential ambiguity with regard to the visual complexity of stimulus images used.

First, all the stimuli were presented one at a time on the screen, with the corresponding names printed under them. Next, participants were shown two pictures at a time with the corresponding name of the image printed under both pictures. (Presentation parameters were matched to those of the actual experiment, except that the same object was pictured in both locations for these familiarization trials.) Finally, to verify that participants correctly identified the pictures, each picture was presented individually at the center of the screen with three words beneath it and participants were instructed to select the word corresponding to the picture.

Experimental Trials

Participants began each experimental trial by clicking on a "START" box at the bottom center of the display. They were encouraged to keep the mouse moving continuously once the cursor entered the start box and instructed to move the cursor upward through a white rectangular box, which cued the audio stimulus to play midflight. Once they heard the spoken word, their task was to make a selection by clicking on one of the two images presented on screen in the top left and top right corners and thus end the trail

If the spoken word they heard was "beaker", their task was to click the target image of the beaker. The distractor image might be a cohort competitor (e.g., "beetle"), a rhyme competitor (e.g., "speaker") or an unrelated control (e.g., "carriage").

Each participant completed an experimental block of 35 trials total, which was composed of 8 cohort competitor trials, 8 rhyme competitor trials, 8 unrelated competitor trials, and 11 filler trials. Thus the eight target words, taken from the eight referent sets used in Allopenna et al. (1998), appeared in each of the three conditions. Filler trials were included in order to prevent participants from developing clues regarding the purpose of the study.

Analysis

Mouse movement trajectory recordings began once participants clicked on the Start button and ended when the target image was clicked. The data collected for each of the trials included the x, y coordinates of the computer mouse along with time in ms. Error trials in which participants had failed to click the target image were excluded from further analyses. All remaining trajectories were visually inspected for cross-overs and obvious sporadic movement (loops, stops, etc.). Such sporadic trajectories were also excluded from further analyses.

All analyzable trajectories were time-normalized to 100 time-steps to account for the potential for trial duration variability following a procedure originally described in Spivey et al. (2005). All trajectories were aligned so that their first observation point corresponded to (0,0) and right-branching trajectories were reflected in the y-axis.

Maximum deviation was calculated as the furthest deviation of the mouse from a straight line connecting the "Start" box to the target image. Reaction time was defined as the amount of time elapsed between clicking the Start button and clicking the target.

Mouse tracking data were analyzed by subject using repeated measures ANOVA with Group (EL1, KL1) as a between-subjects factor and Condition (cohort, rhyme, unrelated control) as a within-subject factor and Subject nested in Group as the error term. Tukey post hoc tests were used to examine pairwise contrasts when significant main effects or interactions were detected. Linear regression analysis was used to explore the relationships between the subject variables (language proficiency, AoA) and the mouse tracking data. A p level of less than 0.05 was considered significant for all analyses.

Results

Accuracy and Reaction Time

Accuracy for both groups was extremely high (greater than 97% for all conditions, see Table 2). Inferential tests for condition or group differences in accuracy were not informative due to perfect performance in half of the cells. Mean reaction times were also similar between groups, but varied significantly by Condition, $F(2,162) = 18.98$, $p < 0.05$. Tukey post-hoc analysis revealed that reaction times were significantly longer for both the cohort condition and the rhyme condition compared to the Unrelated condition. The effect of Group was not significant, nor was there any interaction between Group and Condition.

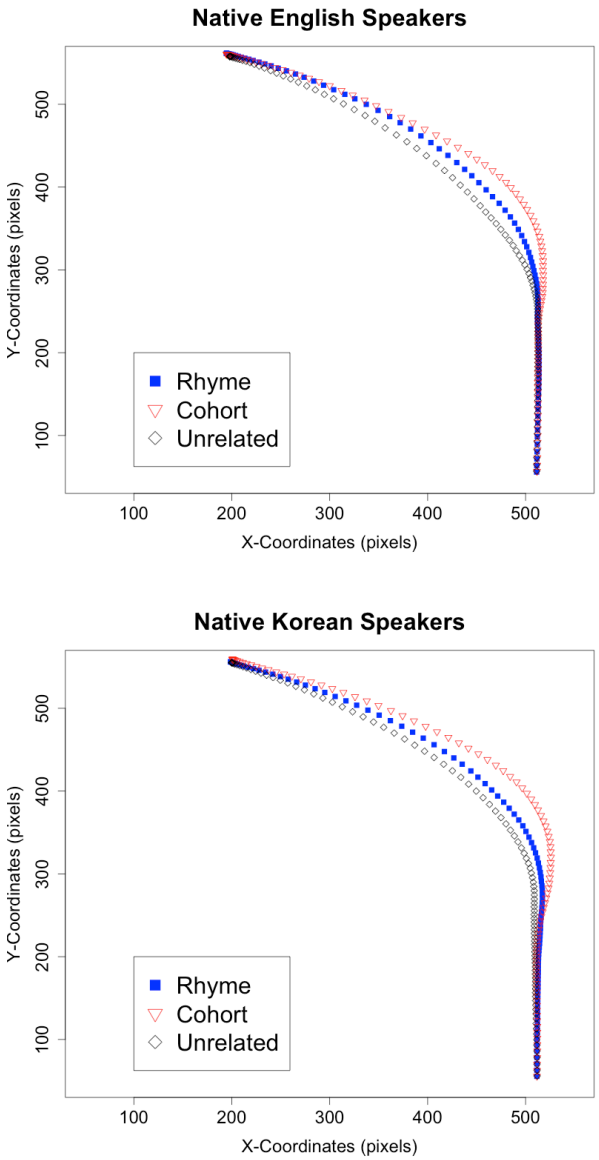

Figure 2. Mean Trajectories for the Rhyme, Cohort and Unrelated Conditions

Table 2: Mean Accuracy & Reaction Times

Group	Cond	Accuracy	RT
EL1	Unrel	1.00 (0.00)	996 (32)
	Coh	0.98 (0.01)	1082 (40)
	Rhy	1.00 (0.00)	1031 (34)
KL1	Unrel	1.00 (0.00)	1002 (32)
	Coh	0.97 (0.01)	1090 (32)
	Rhy	0.99 (0.00)	1053 (33)

Condition means with standard error in parentheses.
RT = Reaction Time (ms); Unrel = Unrelated;
Coh = Cohort; Rhy = Rhyme.

Maximum Deviation

Maximum deviations along with the 95% confidence intervals are plotted by Condition in Figure 3 for the EL1 and KL1 groups. In general, it was observed that large deviations were seen in the cohort condition for both groups. Following ANOVA analysis, a significant main effect of Condition was observed, $F(2,162) = 27.27$, $p < 0.05$. Tukey post-hoc analysis revealed that maximum deviations were significantly larger for the cohort condition than both the unrelated and the rhyme conditions. In addition, the maximum deviation for the rhyme condition was significantly greater than for the unrelated condition. The effect of Group was not significant, nor was there any interaction between Group and Condition (all $F < 1$).

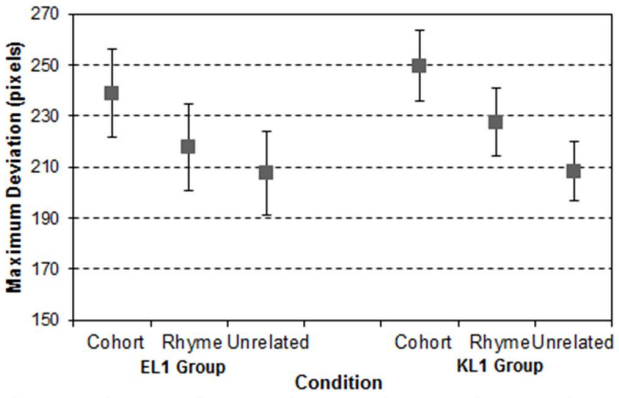

Figure 3. Maximum mouse deviation by Condition for both the EL1 group and KL1 group. Bars indicate 95% confidence intervals.

In light of the above findings, two new difference variables were created to examine relationships between phonetic and vocabulary abilities as well as AoA and LoR with different aspects of performance in the mouse-tracking task. The first was the "cohort effect" which was calculated as the difference between the maximum deviation for the cohort condition and the unrelated condition. The second was the "rhyme effect" which was calculated as the difference between the maximum deviation for the rhyme condition and the unrelated condition. In serial simple linear regression analyses of data from KL1 participants, individual differences in vocabulary, fluency, and phonetic discrimination had no significant correlation with the strength of either the cohort effect or the rhyme effect. AoA and LoR also showed no significant correlation with the cohort or rhyme effect, all $r^2 < 0.01$, all $F < 1$.

Discussion

The most striking feature of the results reported here is the robustness of the cohort and rhyme effects in second-language listeners. Consideration of both the functional architecture of lexical processing, and the strategies deployed to resolve ambiguity led us to predict differences in the relative size of one or both of these effects in listeners with non-native-like speech sound categorization and vocabulary abilities. Nonetheless, the dynamics of lexical activation showed sensitivity to anticipatory information and to overall similarity that was robust to differences between L1 and L2 groups; further, we found no evidence for continuous relationships between measures of English speech sound categorization, vocabulary or verbal fluency and the size of these effects.

The bilingual participants in our study performed well below native levels in a task directly testing their phonetic categorization abilities. Other populations with such difficulties, such as individuals with reading disability (Desroches, et al., 2008) and varying levels of language impairment (McMurray, et al., 2010) have distinct patterns of performance in tasks that tap the lexical-phonetic interface. The current data constrain the interpretation of those previous findings, by demonstrating that differences in the dynamics of lexical activation may not result from atypical activation of appropriate phonetic constituents of words, when this is the result of performing the task in an L2. This is consistent with McMurray et al.'s TRACE simulations, in which differences in lexical decay – rather than activation parameters related to phoneme activation -- were found to best fit the impact of language deficits on relative levels of competition in this same task.

Our sample size was sufficient to explore continuous associations between the dynamics of lexical activation in this task and a range of biographical data and proficiency measures. None of these variables were correlated with the magnitude of the cohort or rhyme effect. Thus, these effects may be robust features of proficient spoken word recognition that do not depend on native-like phonological or lexical processing, or on learning one's second language during a particular sensitive period.

We had also tentatively predicted that the cohort effect would be smaller for non-native listeners based on strategic considerations. If non-native listeners are aware that they experience greater ambiguity as spoken words unfold over time, they may accumulate more information (i.e., wait longer) before committing to one or another response. This waiting strategy may have impacted either reaction time or maximum deviation in the mouse-tracking measure.

There was no evidence for this hypothesis, but it would be premature to rule out strategic differences in task performance entirely. The visual world paradigm is designed to provide data about lexical processing in a communicative context, and to provide evidence for the relative activation of competing words during word recognition. In order to do this, however, it provides a very specific context — especially when adapted to mouse tracking, in which only two choices are presented — that can be strategically used to reduce the inherent vagueness of the spoken stimulus to mere ambiguity between two options that are known in advance. If, as we have speculated, second language listeners are particularly adept at using contextual information to overcome particular difficulties that arise for them due to their non-native phonological and lexical inventories, processing differences may arise in a less direct test of competition among cohort and rhyme neighbors. That is, the similarity between L2 and native listeners in this task may be the result of L2 listeners' ability to take advantage of the limited response set.

Much of the literature on word recognition in second language learners and bilinguals has focused on whether there is competition between lexical representations for different languages, and how this competition is managed (Weber & Cutler, 2004; Marian, Blumenfeld, & Boukrina, 2008). These studies have largely focused on the inter-lingual cohort effect, in which similar-sounding words in L1 appear to be transiently activated, even in a monolingual test run in L2 (Spivey & Marian, 1999).

In this study, we focused on competition effects within a single target language using a mouse-tracking paradigm, and found, quite surprisingly, that the dynamics of lexical activation in a second language are essentially identical to what is found in a population of monolingual native listeners. In a prior study, Canseco-Gonzalez et al (2010) found subtle differences in the timing of the cohort effect between Spanish-English Bilinguals who differed in whether their L1 was Spanish or English. In contrast, we found no difference in the size of the cohort effect between native and non-native English speakers, despite gross differences in proficiency on a number of language measures. Importantly, the group differences observed by Canseco-Gonzalez et al. were based on real time analyses -- the cohort effect emerged later for Spanish L1 than English L1 bilinguals. In our study, there was no difference in a spatial measure of lexical competition -- maximum deviation of the mouse trajectories. This suggests that the *relative* time course of activations for competitors is substantially the same as in L1 speakers, although it is possible that finer-grained analyses of response dynamics might reveal temporal differences in processing. How L2 listeners accomplish native-like performance in spoken word recognition despite lacking native-like speech sound categorization presents a puzzle for models that assume a strictly hierarchical organization — as all current models of spoken word recognition do.

This experiment was designed as a two-alternative forced choice between two similar-sounding options. This task affords different strategies than a "free field" word identification task. Further, the relevant theory (TRACE) is about activation of lexical competitors, and it is possible that the trial preview period artificially activates words that may not otherwise be activated for the L2 speakers before the audio stimuli is presented.

It will be important in future research to study the lexical-phonetic interface under a wider range of task conditions. In particular, exploring the role of the communicative context in shaping L2 listeners' use of anticipatory information. For example, Magnuson et al. (2007) studied the influence of neighborhood structure on the activation of target words in a visual world paradigm in which only unrelated items were present. By examining the activation dynamics of words with different numbers of cohort and rhyme neighbors in this way, they were able to identify effects traceable to patterns of phonetic similarity and lexical activation that may more directly reflect the functional architecture of lexical processing. An adaptation of this paradigm for the current population could include a further manipulation based on lexical overlap in the bilingual participants' first language, to explore the generality of inter-language interference effects observed by, e.g., Marian et al. (2008). Nonetheless, the findings in the current study suggest that, at least under some circumstances, the dynamics of lexical activation are surprisingly native-like in L2 learners, despite gross differences from native listeners in both phonetic inventory and lexicon size.

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 & Language*, 38, 419-439.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. *Language experience in second language speech learning: In honor of James Emil Flege*, 13-34.

Canseco-Gonzalez, E., Brehm, L., Brick, C. A., Brown-Schmidt, S., Fischer, K., & Wagner, K. (2010). Carpet or carcel: The effect of age of acquisition and language mode on bilingual lexical access. *Language and Cognitive Processes*, *25*(5), 669-705.

Desroches, A. (2008). *The temporal dynamics of phonological processing in typical development and in developmental dyslexia*. Ottawa: Library and Archives Canada = Bibliothèque et Archives Canada.

Dijkstra, T., Grainger, J., & Van Heuven, W. J. (1999). Recognition of cognates and interlingual homographs: The neglected role of phonology. *Journal of Memory and language*, *41*(4), 496-518.

Flege, J. E. (1999). Age of learning and second language speech. *Second Language Acquisition and the Critical Period hypothesis*, 101-131.

Kroll, J. F., & Dussias, P. E. (2004). The comprehension of words and sentences in two languages. *The Handbook of Bilingualism*, 169-200.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, *5*(11), 831-843.

Lagrou, E., Hartsuiker, R. J., & Duyck, W. (2011). Knowledge of a second language influences auditory word recognition in the native language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(4), 952.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science*, *31*(1), 133-156.

Marian, V., Blumenfeld, H., & Boukrina, O. (2008). Sensitivity to phonological similarity within and across languages. *Journal of Psycholinguistic Research, 37*, 141–170.

Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human perception and performance*, *15*(3), 576.

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

McMurray, B., Samelson, V. M., Lee, S. H., & Bruce Tomblin, J. (2010). Individual differences in online spoken word recognition: Implications for SLI. *Cognitive psychology*, *60*(1), 1-39.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189-234.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*,*23*, 299-325.

Schulpen, B., Dijkstra, T., Schriefers, H. J., & Hasper, M. (2003). Recognition of interlingual homophones in bilingual auditory word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(6), 1155.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proc. National Academy of Sciences*, 29, 1039310398.

Spivey, M. J., & Marian, V. (1999). Cross talk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological science*, *10*(3), 281-284.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive science*, *34*(3), 434-464.

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory & Language, 50,* 1-25.