How Do Listeners Represent Sociolinguistic Knowledge?

Laura Staum Casasanto (Laura.Casasanto@mpi.nl)
Max Planck Institute for Psycholinguistics
P.O. Box 310, 6500 AH Nijmegen, NL

Abstract

What kinds of representations underlie listeners' knowledge of the way different types of speakers speak (sociophonetic variation)? Listeners store information about the social conditioning of phonetic variation, and use this knowledge to inform their perceptions of speech. But are expectations about sociophonetic variation stored on the level of the word, or do they also apply to never-heard words? A reaction-time experiment investigated whether listeners form the same representations of words for all speakers, or form different representations based on the social characteristics of the speaker. Participants' reactions to real words were compared with reactions to nonce words, in order to investigate the level at which these representations differ. Results showed that social information influenced the processing of ambiguous nonce words the same way it influenced the processing of real words, suggesting that listeners form different representations of speech for speakers with different social characteristics at the level of the sub-lexical 'chunk'. This finding about listeners' knowledge of sociolinguistic variation supports the inclusion of a sub-lexical level of representation in exemplar theories of speech perception.

Keywords: psycholinguistics; speech perception; social.

Introduction

What kinds of representations underlie listeners' knowledge of sociophonetic variation? Listeners store information about the social conditioning of phonetic variation, and use this knowledge to inform their perceptions of speech (Staum Casasanto, 2008). A series of experiments investigated whether listeners have knowledge about *t/d* deletion, a sociolinguistic variable, and, if so, whether social information that listeners gather from the non-linguistic context is used in formulating expectations about sentence meanings. Results indicate that listeners have implicit knowledge about the social correlates of *t/d* deletion, and that they use this knowledge, combined with social information from the scene, in resolving lexical ambiguity, suggesting that social information is a part of language understanding. Information about speakers must be included somehow in listeners' mental representations of linguistic forms. But what is the nature of these representations?

This paper presents results of an experiment addressing the nature of listeners' representations of sociophonetic variation. This experiment contrasts listeners' reactions to real words with their reactions to nonce words, in order to investigate whether listeners' knowledge of social constraints on phonetic variation is part of lexical representations, or is associated with a level below the word.

Effects of context on speech perception have commonly been accounted for by exemplar models, in which listeners have detailed episodic memory traces of linguistic experiences that include details not only of the acoustic signal they perceived, but of many aspects of the context in which the signal was perceived (Goldinger, 1996; Johnson, 1997; Pierrehumbert, 2001). The original motivation for importing such models from perception and categorization in general to speech perception was to account for detailed phonetic knowledge that speakers and listeners have about specific words in their lexicons; exemplar models can account for the effects of lexical frequency on phonetic reduction, for example (Bybee, 2000). However, these models can also account for effects on the level of groups of speakers (and have been invoked to do this as well [e.g., Hay, Warren, & Drager, 2006]). In exemplar models of phonological knowledge, social information about speakers can be available to the listener by virtue of indexing of the tokens of past experiences that are stored by the listener.

Including social indexing in an exemplar model requires that the detailed traces of linguistic experiences include information about the speaker. The stored details of these experiences allow listeners to associate aspects of linguistic form with characteristics of speakers.

Relationships between social characteristics and sociolinguistic variables are thus generalizations across stored tokens. However, the architecture of an exemplar model provides potential limits to the types of inferences that listeners might be able to make about future speaker behavior. Specifically, an incoming token must correspond to a previous token in some way in order to activate details of the previous token as it was experienced by the listener. This aspect of the model makes a testable prediction: social information should only influence the perception of tokens that correspond to previously experienced types.

But what constitutes a type in exemplar theory? In a strict version of exemplar theory, where tokens are episodic traces of previously experienced exemplars, the basic unit of exemplar storage and the abstractions that can be made over these basic units are of crucial importance in making behavioral predictions based on the model. In fact, the potential flexibility of the process of abstraction across the space of stored tokens in an exemplar model can make it difficult to determine at what level the exemplars are actually stored.

While there is no universal consensus on this matter, Johnson (2005) has suggested that the word may be the unit at which exemplars are stored, because words are more accessible to speaker/listeners than sounds. This suggestion allows for a specific behavioral prediction: if social information is included in the model by indexing stored tokens of words with details about the social context in which they were uttered, then sociolinguistic knowledge should apply only to words listeners have already heard.

This experiment investigated whether listeners can apply their knowledge of how different speakers use linguistic variation to the recognition of words they have never heard before. If sociolinguistic knowledge is part of lexical representations, then listeners' perceptions of nonce words should not be influenced by the social characteristics of the purported speaker, because they do not have pre-existing lexical representations of these nonce words.

Methods

Participants

Thirty-nine native English speakers from the Stanford University community participated in this study in exchange for payment. All participants had lived in the United States for at least 18 years. Participants were of a range of races/ethnicities and both genders, and most were between 18 and 22 years old.

Design

The design included four factors with two levels each: Word Type (real vs. nonce), Coda Type (t vs. no t), Face Type (Black vs. White), and Voice Type (Black vs. White).

Auditory Materials

Target items were 24 sentence beginnings (without endings) each containing a target word. Target words in the real word condition were pairs of words that were ambiguous between a word ending in a consonant cluster with a *t* or *d* in the final position (such as *mast*, in the *t* condition) and a word that is identical save for the absence of the *t* or *d* (such as *mass*, in the no *t* condition):

> The *mass* probably lasted…
> The *mast* probably lasted…

The sentence frames were constructed so as to maintain the ambiguity between the two words (the content of the sentence frames was consistent with both interpretations). In the nonce word condition, these target words were replaced by similar sounding word-like strings that were invented for this experiment. Like the ambiguous real words, these nonce words were "ambiguous" between a non-word with a deleted *t* or *d* at the end of a final consonant cluster, and a non-word that is identical except for the final stop in the consonant cluster. These nonce words all had a similar structure to that of the real words, with consonant clusters that could be subject to *t/d* deletion, paired with the words that would be ambiguous with them after deletion (e.g. *stip*/*stipt* or *cliss*/*clist*). Each nonce word was paired with one of the ambiguous sentence beginnings, creating a phrase identical to one of the real phrases except for the nonce word replacing the target word:

> The *frass* probably lasted…
> The *frast* probably lasted…

Each participant heard each carrier phrase twice – once with a real word, and once with a nonce word. These two instances of the carrier phrase were always in different blocks of the experiment. The sound files were excerpted from recordings of entire sentences read by naïve Stanford graduate and undergraduate students, who were paid for their time. Participants heard excerpts from sentences that never contained an underlying final stop (i.e., they heard sentences in which speakers intended to say *mass* but never sentences in which speakers intended to say *mast*). Thus, participants never heard any version of the experimental sentences that contained an underlying *t/d*, so there were no cues in the speech stream to the presence of a deleted stop.

Each target item was heard spoken by an African American speaker by half the subjects and spoken by a European American speaker by the other half of the subjects; the race of the actual speaker (Voice Type) was crossed with the race of the pictured speaker (Face Type). Having both types of voices in the Black and the White Face Type conditions prevented one face condition from being generally more felicitous with the voices heard than the other. However, the acoustic cues to race/ethnicity (other than *t/d* deletion) available in each clip varied naturally, and were not controlled. Listeners could potentially have been influenced by cues to the race of the speaker that were present in the audio clip; in analysis, the actual race of the speakers was used as a proxy for cues to race in the speech stream. If cues from the speech stream are strong enough to influence listeners' reactions, they should do so in the same way that cues from the pictures are predicted to do.

Forty-eight similar fillers were constructed that also consisted of only the beginning portion of a sentence, but did not contain any *t/d*-ambiguous words. One word was selected from each sentence beginning to serve as the false target. In addition, sixteen similarly structured fillers were created that contained words that could be subject to *t/d* deletion without creating ambiguity. For example, the word *fast*, when subject to *t/d* deletion, becomes [fæs], which is not a word in English. These sentences were recorded by a non-naïve speaker[1], who was instructed to produce the words without a final stop. As with the first 48 fillers, the beginning portions of these sentences were used, with the words with deleted final stops serving as false targets. The purpose of these fillers was to make the overall tone of the

[1] This speaker had to be aware of the focus of the experiment because he needed to specifically avoid producing audible final consonants in the ending clusters of the crucial words. These productions may have contained cues to an underlying /t/, but this would not interfere with their function of giving participants reason to believe that *t/d* deletion was compatible with the speech situation of the speakers in the experiment.

experiment more casual, and to encourage participants to believe that the speech they were hearing might contain informal variants like deleted *t/d*. However, these sentences were produced by a different speaker so that participants did not have any a priori reason to believe that a particular speaker of the target words would or would not engage in *t/d* deletion. In total, there were 24 target items and 64 fillers; 40 of these could have been interpreted as containing a *t/d* deletion.

Visual Materials

Each sound clip was presented with a photo of a purported speaker; target items were matched with four pictures of Black males and four pictures of White males. Ordinary filler items (which were spoken by females) were matched with eight pictures of females of various races/ethnicities. One picture of a male who was of East Asian descent was matched with the voice that produced the 16 fillers containing unambiguous *t/d* deletions. All photos were taken from a database of university ID photos from a different university than the participants attended. In all, there were 9 photos of males and 8 photos of females used in this experiment.

Procedure

Participants were instructed to listen to a short sound clip while looking at a picture of a face, which they were told represented the speaker of the clip. They heard the ambiguous portion of one of the sentences, which contained no final stops, e.g., *The [mæs] probably lasted*. While they were listening to the clip, participants saw the words in the phrase they were hearing below the picture of the speaker, with one of the words replaced by an underlined space. This phrase appeared at the beginning of the trial, at the same time that the clip began. Participants then saw either the *t* version or the non-*t* version of each word appear below the picture of the speaker; the word appeared after the clip was finished playing, so that participants had already finished processing the auditory stimuli by the time the written word appeared.

Participants pressed Y to indicate that they believed the word on the screen was the word they had heard (that went in the blank), and N to indicate that they believed the word on the screen was not the word that they had heard. Response times were measured from the time the target word appeared on the screen. In approximately half of the trials (including both targets and fillers) participants were presented with plausible transcriptions of the word in the audio clip, and in the other half they were presented with implausible transcripts of this word, although the target items were all presented with a plausible transcription of the target words.

Each participant either saw Black faces matched with non-*t* words and White faces matched with *t*-words, or Black faces matched with *t*-words and White faces matched with non-*t* words, creating a between-subjects design. Each voice was presented in half the trials paired with one Black face and in the other half of the trials paired with one White face (between subjects), so that the race of the speaker and of the person pictured were crossed. Each subject heard each voice paired with only one picture, to increase the likelihood that the participants interpreted the people pictured as the speakers of the clips.

Results

In the real word condition, when the picture indicated the speaker was Black, listeners responded faster to non-t words (e.g. *mass*) than to t words (e.g. *mast*); when it indicated the speaker was White, this difference disappeared (F1(1,74)=2.32, p(rep)=.86, F2(1,22)=4.56, p(rep)=.93)[2] – see Figure 1). These results reflect the way Black and White speakers tend to produce *t*-words. This effect was produced despite the fact that listeners actually heard the same acoustic input in all cases, suggesting that the difference observed was in the way listeners categorized that acoustic input on the phonetic level.

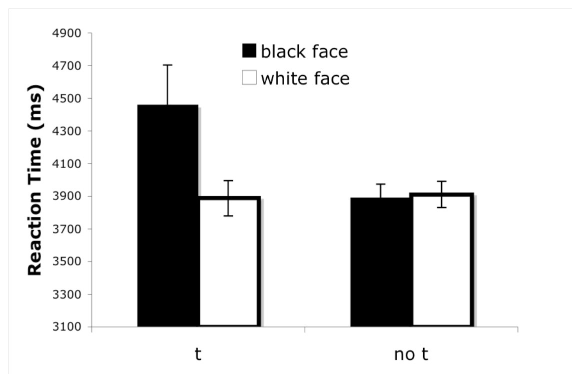

Figure 1: Results of real word condition. Participants responded faster to *t* words when they saw a White speaker than a Black speaker, but not for non-*t* words. Error bars represent s.e.ms.

The actual race of the speakers of the clips also influenced reaction times, independent of the race information from the pictures. Listeners responded faster to the non-*t* words when the actual speaker of the clip was African American than when he was European American, and they responded faster to the *t*-word when the speaker of the clip was European American than when he was African American (F1(1,74)=2.9, p(rep)=.88, F2(1,22)=9.38, p(rep)=.97), consistent with the results based on the race of the pictured purported speakers.

[2] P-rep indicates the probability of replicating an observed effect, given an equipotent replication (Killeen, 2005). A p-rep value of .92 corresponds to a 2-tailed p-value of .05, and can be interpreted as estimating a 92% probability of a replication producing a difference with the sign in the same direction as the observed difference.

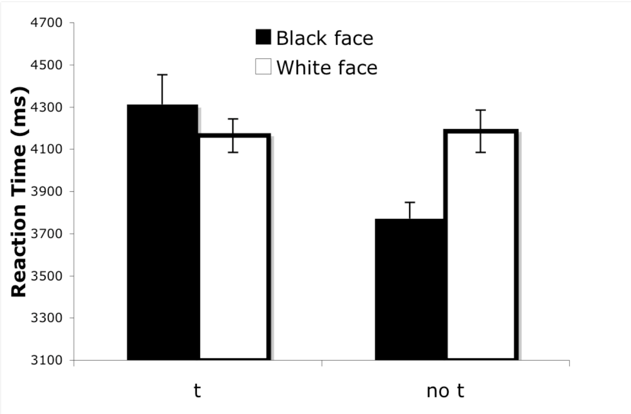

Figure 2: Results of nonce word condition. Participants showed the same interaction in reaction times when responding to nonce words as when responding to real words. Error bars represent s.e.ms.

Similarly, in the nonce word condition, when listeners believed the speaker was Black, they formed more t-less representations (they responded faster to *frass* than to *frast*) than when they believed the speaker was White (when this difference disappeared) (F1(1,74)=3.01, p(rep)=.88, F2(1,20)=32.24, p(rep)=.99 – see Figure 2).

Also parallel to the real word condition, the actual race of the speaker whose voice listeners heard influenced reaction times. Nonce words with a final stop were responded to faster when the speaker of the clip was European American than when he was African American, and nonce words with no final stop were responded to faster when the speaker was African American than when he was European American (F1(1,74)=2.95, p(rep)=.88, F2(1,20)=23.23, p(rep)=.99). Importantly, the interaction between Face Type and Word Type in nonce words was not significantly different from the interaction found in real words (alls Fs<1). The interaction between Voice Type and Word Type also did not differ from the interaction found in the real word condition (all Fs<1). The fact that these effects of race information on speech perception persisted in the nonce word condition and did not differ statistically from the effects found in the real word condition indicates that social information influences word recognition the same way for words that have pre-existing lexical representations and words that do not.

Discussion

Results in the real word condition demonstrated that listeners form different phonetic representations of words that could have been subject to *t/d* deletion for speakers of different races. Results in the nonce word condition demonstrated that the influence of social information on speech perception extends to novel words. If the unit of exemplar storage is the word, then listeners should not make inferences about how different speakers would pronounce words for which they do not yet have a lexical entry, and of which they have not yet stored any traces. Contrary to the prediction of a word-based exemplar theory, social information influenced word recognition just as much in the nonce word condition as in the real world condition. These results suggest that social information must be accessible to processing at some other level(s) of representation. While units of sound such as phonemes or phones may seem to be an obvious alternative, the fact that the variable phenomenon in question involves a fully deleted variant makes this alternative unsatisfactory.

A reasonable phonetics-level account can be constructed of the observed results in the *t*-word condition. Seeing a White face, according to an exemplar model, may activate to some extent all utterances ever made by a White speaker in the listener's experience (because these utterances are "indexed" according to contextual factors, of which speaker race is one). Assuming that Black and White speakers produce approximately the same number of phonemic /t/s, the fact that White speakers do less *t/d* deletion suggests that they on average produce more phonetic [t]s than Black speakers. Thus, the phonetic [t] activated by the orthography of the word *mast* may be more consistent with a White speaker, in this model, simply because White speakers produce more [t]s in general. This mechanism could result in the observed differences between responses to the Black and White face conditions for the *t*-words in both the real word condition and the nonce word condition.

However, while a phonetics-level explanation could account for social influences on participants' responses to seeing a *t*-word, this explanation does not seem to apply to the cases where no orthographic *t* was seen. When the word *mass* is seen, it activates the phonetic representation [mæs], which matches more exemplars of words spoken by Black speakers, because not only their productions of *mass* match this, but also more of their productions of *mast* – a lexical effect.

It is also presumably true that seeing a non-*t* word (such as *mass*) activates representations of the individual phones that correspond to the letters in the word's orthographic representation (in the case of *mass*, this would be [m], [æ], and [s]). However, these activations would be no different based on the race of the speaker, because neither Black nor White speakers habitually delete these phones. The lack of a difference in production by these speakers suggests that there should be no difference in the number of previously experienced tokens available to be activated in the mind of the listener.

While it may also be the case that, parallel to the explanation of the *t*-word effect above, overall Black speakers produce more words without a [t], it is hard to see how the lack of [t] could be represented in a way that would spread activation to other words that do not have a [t]. Thus, this phonetic representation is more consistent with a Black speaker, but only because of factors related to tokens of these specific words – not because of general tendencies over exemplars of phonemic /t/ or phonetic [t]. The phonetic account predicts that the differences between Black and White speakers should be restricted to *t*-words in the nonce word condition, which is not consistent with the results, which show a significant difference between the race conditions in the non-*t* words (t1(35)= 2.09, p(rep)=.93; t2(22)=5.3, p(rep)=.99). Thus, a phone- or phoneme-level explanation fails to account for the nonce word results.

The finding that an exemplar theoretic account of this effect cannot depend on words or on phonemes raises a question: how *do* listeners represent correlations between social characteristics of speakers and linguistic variation?

Even though participants have likely never heard the word *frass* before, they are probably able to recognize parts of that word as similar to words they have heard before. Even though correspondences on the level of the phone cannot account for their behavior, an exemplar account of this effect could potentially be constructed if it is possible for incoming tokens to match up to previously experienced tokens at some intermediate level, like sublexical chunks of phones. The end of the word *frass* does correspond to many of previously experienced tokens of other word *endings,* and because of their different rates of *t/d* deletion, it matches up to more word endings spoken by Black speakers than by White speakers.

Thus, it is not impossible to create an exemplar-based explanation of this effect; however, the current results constrain what kinds of correspondences will be necessary to account for listeners' behavior. Including representations of sublexical chunks of phonetic material in a model of speech perception allows social information to influence the perception of never-before-heard words, as long as these words contain some previously heard sublexical chunks.

If an exemplar model of speech perception is to account for the results of this experiment, it will need to include abstractions not only at the levels of the word and phoneme, but also at the levels of the phone and the sub-lexical chunk. In the case of an exemplar model, these abstractions can emerge from patterns in listeners' detailed episodic traces of speech events. However, not all types of models of speech perception would require the reification of sublexical chunks of phonetic material to account for the results of the current experiment. For example, listeners' perceptions could be influenced by social information about speakers if they mentally model the speaker's grammar during perception (Flemming, 2009). If listeners make inferences about constraint rankings or application of a variable rule (two alternative ways of modeling variability in the realization of *t/d*), only phonetic and phonological representations are necessary to predict the results of the current experiment. Determining which levels of representation are involved in making the inferences depends on the mechanisms at work in the listener's mind.

If phonetic categories and sub-lexical chunks are input to the computation underlying inferences that listeners make about speech based on characteristics of speakers, then these categories, as well as the more well-established lexical and phonological categories, are necessary for speech perception. These abstractions would allow listeners to associate incoming tokens with types on a variety of levels, so that new tokens that correspond to previously experienced tokens on some levels but not others can be accurately categorized. It appears that social information about the speaker can help to determine what constitutes a correspondence between an incoming token and previously experienced tokens.

In any contextually sensitive model of speech perception, a nearly limitless supply of information could be used to constrain the categorization of incoming tokens. In theory, listeners could have stored details about the time of day of each utterance they hear, or the color of the walls in the room they were in when they heard it. However, many of these types of information would not be helpful to the listener in his or her quest to correctly interpret speech.

Ideally, the listener would have some means of evaluating the informativity of the details of the situation accompanying the speech he or she experiences, and would be biased to use informative details to constrain their categorizations of incoming tokens, while ignoring uninformative details. A Bayesian model of language comprehension (such as that described in Norris and McQueen, 2008) builds this useful bias into the evaluation of the listener's hypothesis about what the speaker is saying. This type of model includes both a parameter that represents the conditional probability of an event given a contextual factor, and a parameter representing the conditional probability of the alternative given the same contextual factor. Only if the conditional probability of the event is different from the conditional probability of its alternative does the existence of the contextual factor influence the estimate of the probability of the event. Thus, any contextual factors that are not informative with respect to the event will have no effect on such a system.

In a model of language understanding as Bayesian inference, determining what word has been uttered is equivalent to assigning a probability to an interpretation of the speech stream. Listeners use their knowledge of the relationships between social information and linguistic variation in evaluating the hypothesis that a speaker would delete a *t* or *d*, given the race of the speaker (socially influenced speech perception). How can the listener use information about the speaker's race to assign a probability to a deletion event?

It is impossible to directly query the probability listeners have assigned to the *t/d* deletion, which is what the Bayesian model makes a prediction about. However, there are behavioral correlates of this probability that can be measured. If listeners assign a higher probability of retaining the *t* in the word *mast* to White speakers, then seeing an orthographic representation of this word, which activates the phonetic representation of a [t], will be most consistent with representations formed when the listener believes that the speaker is White. In this experiment, reaction times to *t*-words like *mast* were faster when the pictured speaker was White, consistent with the predictions of the Bayesian model.

The proposal that using social information can help listeners interpret the speech stream provides a functional motivation for listeners to use social information in speech perception. However, in the experiment presented in this paper, the use of information about the race of the speaker sometimes caused listeners to impute a *t* that was never actually present in the speech signal. This suggests that social information is not inherently helpful – being influenced by information about speaker race could potentially be detrimental to the listener. This raises the question, why would we have a speech perception system in which social information sometimes causes us to hear things that weren't there?

The answer lies in the difference between the conditions that exist in the laboratory and those that exist in the natural world. While the human speech perception system may not be perfectly adapted for the tasks listeners performed in these experiments, the lab differs crucially from real life in that social information and linguistic behavior were varied independently from one another in these experiments. In the real world, social characteristics and phonetic realizations tend to covary, which is exactly how listeners develop these different estimates of the parameters in the equation in the first place. It is still possible for social information to steer a listener in the wrong direction in the real world; however, as long as social factors and linguistic behavior are statistically correlated, the social information is helpful, on average.

In the simplest Bayesian framework for understanding inferences from context to speech, listeners have perfect information about the context, and imperfect information about speech. However, in real life, listeners rarely, if ever, have perfect information about anything. In many situations, listeners may be using their beliefs about a speaker's use of *t/d* deletion to make inferences about their race *and* using their beliefs about the speaker's race to make inferences about their use of *t/d* deletion at the same time. In such a situation, the Bayesian model makes a prediction: hearing tokens of *t/d* deletion early in an encounter should make later, ambiguous tokens more likely to be interpreted as deletions. When the listener hears a token and classifies it as deleted, this increases the likelihood they assign to the speaker being Black. Because the output of this process is the input to the process of socially influenced speech perception, this in turn makes all following tokens more likely to be interpreted as deleted tokens. Thus, when the listener has imperfect information, the act of classifying tokens and categorizing the speaker changes the way the listener classifies tokens and categorizes the speaker in the future; a Bayesian model of this process predicts perceptual learning (Goldstone, 1998) of these categories.

Conclusions

Listeners use social information to inform their perception of speech, both when understanding real words and when understanding novel words. This suggests a role for phonetic representations in any model of speech perception. Furthermore, a successful exemplar model of speech perception will have to allow for abstraction over sublexical chunks of phonetic material. More generally, these results suggest that a rational approach to speech perception may provide insight into the types of information listeners mentally represent and the ways in which those types of information interact to produce the expectations and inferences that underpin our understanding of language in real time.

Acknowledgments

Thanks to D. Casasanto, H. Clark, K. Drager, P. Eckert, E. Flemming, R. Greene, L. Hall-Lew, S. Lewis, B. Munson, J. Rickford, R. Starr, M. Sumner, and A. Zwicky for helpful discussion. This research was supported in part by a Mellon Dissertation Year Fellowship to the author.

References

Bybee, J. (2000). The Phonology of the Lexicon: Evidence from Lexical Diffusion. In M. Barlow and S. Kemmer (eds.) *Usage-Based Models of Language*. Stanford: CSLI Publications.

Flemming, E. (2009). Commentary: The Role of Grammar in Adaptation and Imitation. Presented in the special session on The Culture-Phonology Interface at the 2009 LSA Annual Meeting in San Francisco, CA.

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

Goldstone, R. (1998). Perceptual Learning. *Annual Review of Psychology*, 49 (1), 585-612.

Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics, 34* (4), 458-484.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson, & J. Mullenix, *Talker Variability in Speech Processing* (pp. 145-166). San Diego: Academic Press.

Johnson, K. (2005). Decisions and mechanisms in exemplar-based phonology. *UC Berkeley Phonology Lab Annual Report*, 289-311.

Killeen, P. R. (2005) An alternative to null hypothesis significance tests. *Psychological Science*, 16, 345-353.

Norris, D., & McQueen, J. (2008). Shortlist B: A Bayesian Model of Continuous Speech Recognition. *Psychological Review, 115* (2), 357-395.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee, & P. Hopper, *Frequency effects and emergent grammar*. Amsterdam: John Benjamins.

Staum Casasanto, L. (2008). Does Social Information Influence Sentence Processing? *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, Washington, D.C.