Phonetic Sensitivity in Adult Word Learning

Lisa Cox

Department of Linguistics

Northwestern University

Abstract

What level of acoustic detail do learners use during early stages of word learning, and can changing the acoustic properties of a new word improve learning? We examined these issues in a pre-registered partial replication of White, Yee, Blumstein, and Morgan (2013). Listeners were taught non-words as labels for images, with 1, 5, or 8 training trials per word. At test, participants simultaneously saw two images on the screen, one familiar from training and one new, and selected the image that best matched an auditorily-presented non-word. This non-word was either the trained target, a single-feature mispronunciation of that target, or a novel non-word. Similar to White et al. (2013), we find good performance on trained targets and novel non-words, but poor performance on mispronunciations. This suggests word learning initially makes use of a broad, underspecified phonetic representation that encompasses the target word and near phonetic neighbors. Experiment 2 changes the nature of the auditory input to include multiple tokens of the same word, from the same talker. This additional acoustic variability did not specifically improve learners' rejections of mispronounced words; even in the presence of acoustic variability, word learning makes use of phonetic representations that encompass near phonetic neighbors.

Introduction: Variability and Distraction

How do language learners recognize all, and only, the appropriate acoustic strings as instances of a given word? Learning a new word – in a new language, or even in a native language – can be tough. The new words we encounter as adults are often complex words, jargon specific to a field, or otherwise low frequency. And yet humans are remarkably successful at learning words, with the average adult monolingual's vocabulary estimated to be between 10,000 and 100,000 words (Milton & Treffers-Daller, 2013). Once children start learning words, they are prodigious word learners; between 12 months and 24 months, children typically experience rapid vocabulary growth and can go from knowing no words to 300 (Ganger & Brent, 2004), and afterwards can learn 10 to 20 new words a week (Berk, 2003).

In the section that follows, we will consider the problem generally before turning to specific findings. As hard as it might be to remember that the Spanish word for *duck* is *pato*, or that *electrocardiogram* refers to recordings of electrical signals from the heart, it is a much harder problem if you consider the data we work with. Word learning relies on incredibly sparse and highly variable data on several levels. We create robust, accurate lexical representations, so our cognitive word learning machinery must be fairly sophisticated for humans to ever learn a word. It must be designed to not just 'make do' with variability, but also make *use* of that variability.

When you read the word "cupcake", you know that this is a label with a referent: a small, delicious pastry. You know cupcakes often have icing, come in many flavors, and are usually considered a dessert. You could point to a cupcake in the world, if asked, and if you see a cupcake you can say, "yes, that's a cupcake!" As a speaker, or at least reader, of English, you know that there are a set of sounds that make up this word that (roughly) correspond to letters, and you could probably break the word into those component sounds if pressed – the 'k', 'uh', 'p', etc. You also know that changing one of those sounds changes the word. "Cupca**p**e" no longer means a small, delicious pastry; it doesn't mean much of anything. For other words, like as "cake" to "ca**p**e", such a change to one letter (or sound) changes the meaning entirely.

Now, when you *hear* the word, you also get a whole world of acoustic information. The word "cupcake" said by ten different people is going to sound slightly different every time – faster or slower, higher pitched or lower, spoken clearly or garbled over a bad phone connection, syntactic and prosodic position in the sentence, accented speech, and many other factors can influence the acoustic realization. Even for an individual speaker, no two acoustic tokens of the same word are the same. Sounds, even within a word, vary as well. The three /k/s in "cupcake" are going to be different, both because of random, slight differences in articulation, and because of systematic, language- and speaker-specific parameters. For example, word-initial /k/ is often accompanied by aspiration in English, while word-final /k/ is likely to be unreleased. Despite all of this variability, present in every token of "cupcake" you will ever hear, you can recognize that every one of those tokens is the same word, the same label picking out the same kind of delicious object in the world.

This ability to form categories, at both the word and sound level, will be referred to as generalization or abstraction. Going from acoustic detail to abstract category representations allows us to generalize our knowledge from specific instances and make use of it in the future. While some of this variability is random, much is systematic, and it is because of that systematicity that we can abstract from the variable signal and categorize a particular acoustic production. While it may seem simpler to have a model of word learning that dismisses acoustic detail and represents words as sets of abstract sound categories, we will argue that tracking – and storing – low-level acoustic variation is key to forming abstract representations at an appropriate level. You need many tokens, in many environments, of /k/ to form a category /k/, and similarly you need many tokens, in many environments, to recognize all and only instances of "cupcake" as being "cupcake".

In this paper, we review prior experimental and modeling findings on the role of acoustic variability in numerous perception and learning situations. From these findings, we hypothesize that learning generalizable speech categories relies on acoustic variability. We discuss one situation, the Switch Task (Stager & Werker, 1997), in which both infants and adults (White, Yee, Blumstein, & Morgan, 2013) make surprising categorization errors during word learning. We then present two experiments: Experiment 1 replicates the task from White et al. (2013), and Experiment 2 modifies the task to incorporate acoustic variability from a single talker (not present in prior versions of the Switch Task). We predict that learners will make fewer categorization mistakes in Experiment 2 than Experiment 1, due to this added variability. Experiment 1 reproduces the expected finding, and Experiment 2 shows that single-talker acoustic variability only improves task performance for specific conditions. We discuss these results in the context of recent work that emphasizes the importance of input from multiple talkers during learning.

Acoustic Variability: Benefits and Challenges

Word Learning and Recognition

There is a wide literature showing that variability is important for categorization and generalization in children's word learning. Much of this has focused on variability in the objects being labeled (e.g. Gentner & Namy, 1999; Liu, Golinkoff, & Sak, 2001; Waxman, 2003). In 2-year-olds, seeing multiple examples of an object with a new word improves word learning, and they benefitted more from a narrow range of variability than a wide range (Twomey, Ranson, & Horst, 2014). Seeing pairs of different items promotes appropriate (limited) categorization, while pairs of identical items do not (Kovack-Lesh & Oakes, 2007). Even variability in the visual background and not the objects themselves improves word learning in 2-year-olds, showing that variability not relevant to the category can be beneficial (Twomey, Ma, & Westermann, 2018). In contrast, during adult L2 learning, varying the object impedes learning (Sommers & Barcroft, 2013).

Both infant and adult learners are sensitive to *acoustic* variability in word forms and labels, as well. Infants at 7.5 months old can recognize a new word when it is presented in a

different amplitude, but it takes until 9 months for infants to recognize the same word presented with a different pitch (Singh, White, and Morgan, 2008). For adults, talker variability can make recalling a list of known words more difficult (Mullenix, Pisoni, & Martin, 1989).

Not all effects of variability are negative. Goldinger and Azuma (2004) found that in a word-reading task, speakers' productions after listening to recordings of another speaker were judged by independent evaluators to be more similar to that speaker than productions made before listening. This means that they must have retained acoustic information about the word productions they heard, and that information influenced their word representations, and therefore their own future productions. Listeners improve at word recognition tasks with increased familiarity with a speaker (Nygaard & Pisoni, 1998), while varying talker characteristics can improve or hinder word recall, depending on the features varied and the time scale of the test (Bradlow, Nygaard, & Pisoni, 1999; Nygaard, Sommers, & Pisoni, 1995). Variability is also beneficial when learning new words. Barcroft and Sommers (2005) showed that multi-talker training on second-language word lists improves word learning, while hearing the same speaker using multiple speech styles did not.

In sum, listeners track and use the acoustic, visual, and social variability of speakers, and make use of that information during word recognition and learning (see Foulkes & Hay, 2015 for a review of sociophonetic structure; see Kleinschmidt, 2018 for a broad review of findings in acoustic and social talker variability).

Sound Category Learning and Recognition

Variability is also useful during sound category learning. Bradlow, Pisoni, Akahane-Yamada, and Tohkura (1997) trained Japanese learners of English on perception of the /r/-/l/ contrast. The learners identified the word from minimal pairs using /l/ and /r/, with tokens from multiple speakers of English. Not only did this improve their perception of the /r/-/l/ contrast, it also improved their production, as evaluated by a separate group of listeners. Having fewer interlocutors makes you more likely to change your boundary between /t/ and /d/ (Lev Ari, 2017), indicating that sound category representations are sensitive to talker specificity.

It is clear that acoustic variability is both a problem and a tool for learners, but what aspects of variability are helpful and which are obstacles seems to be dependent on the learner's task and characteristics. Given that variability in the acoustic signal is retained and used at both the sound and word level, we have two questions: why is it useful, and what kinds are useful for word learning?

Modeling Variability

We propose that acoustic variability is useful because it provides learners information about the parameters that govern the distribution of sound categories in the acoustic space, in terms of both 1) which acoustic dimensions are important (for early learners who have not yet established the set of possible sound categories) and 2) how sound categories are distributed along the relevant dimensions.

Historically, sensitivity to variability has been modeled using associative learning and connectionist approaches (for a review, see Apfelbaum & McMurray, 2011). This can also be understood with a Bayesian inductive inference approach. For a Bayesian listener, the act of perceiving a single speech sound involves hearing an acoustic production, *s*, and inferring the category of the sound the speaker intended to produce, *c*. This category *c* is a distribution that describes how likely each value is to occur for a given set of possible values (Pajak, Fine, Kleinschmidt, and Jaeger, 2016). The probability distribution is defined by mathematical parameters that describe the possible values for each acoustic feature of the category (e.g., for sounds like /p/ and /b/, voice onset time or VOT) [1]. The speaker draws a sample from the distribution of *c* to produce *s*. Other factors may systematically, or randomly, interact with *c* to produce the acoustic realization *s*, including noise, sentential or lexical position, speaker identity and characteristics, accentedness, and many others. For example, plotting a histogram of VOTs of tokens of all bilabial plosives in English would reveal a bimodal distribution, with one peak centered on the mean VOT for /b/ and another centered on /p/. This means that when drawing from the distribution of VOTs for English bilabial plosives, it is likely that a speaker would produce something close to the mean values of either /b/ or /p/ - but for any number of reasons might produce an atypical token from the far edges of the distribution. The listener only has access to *s*, and must infer category *c*. All this to simply perceive a single speech sound!

Sound category learning can be formalized as a hierarchical Bayesian model in which both the category from which *s* is drawn and the categories' parameters must be inferred. Both infant and adult second language (L2) learners do not know what categories exist in a given language, the relevant acoustic dimensions for those categories, or even how many categories there are. For example, English learners need to know there are two bilabial plosive categories /p/ and /b/, as well as the means and standard deviations of the defining features, such as voice onset time (VOT) and aspiration, of those categories. Thai learners, on the other hand, need to learn that there are three bilabial plosives, /p/, /b/, and /p^h/ (Tingsabadh & Abramson, 1993), as well as the parameters governing the distribution of those plosives.

Bayesian models of learning have shown that word knowledge can aid sound category learning, and that this can happen during simultaneous word and sound category learning – precisely the situation that infants, and second-language-learning adults, face. A model given a randomly simulated set of vowels, with values based on the means and standard deviations from the Hillenbrand corpus (Hillenbrand, Getty, Clark, and Wheeler, 1995), and tasked with sorting those vowels into categories did markedly better when the model simultaneously had to learn the words those vowels were in (Feldman, Griffiths, Goldwater, and Morgan, 2013a). Infants also learn vowel categories better when those vowels are embedded in words (Feldman, Griffiths, Goldwater, and Morgan, 2013b). In other words, these models perform similarly to infants, in that both benefit from learning sound categories (with representative acoustic variability) while learning words.

[1] This can be a high-dimensional distribution, considering every possible feature for which *c* can take on a value, but here we will consider only one dimension for simplicity.

Similar models can also account for generalization – taking learning from one context and applying it, selectively, to other contexts where it is useful. Thompson and deBoer (2017) asked whether a Bayesian model can transfer learning about the shape of the distribution of a feature like voice onset time from one place of articulation appropriately to other places of articulation. Importantly, for their task, this 'appropriateness' required generalizing learning (e.g. the number of sound categories) *only* to the other places of articulation that also have the same number of categories. They found that when such a model is given a parameter that is allowed to be updated with learning, it can appropriately generalize only to those categories for which generalization would be helpful.

Pajak, Creel, and Levy (2016) tested two groups of L1-Korean and L1-Mandarin speakers (all bilingual, L2 English speakers) on a perceptual discrimination task and a word learning task. The words in the tasks were either dissimilar, e.g. differing in several phonemes, [tala]-[kenna]; similar, in which the words differed by one phoneme where the difference was along a familiar phonetic dimension for both groups (e.g. [tala]-[taja]; or highly similar, in which the words differed by one phoneme, and the phonetic dimension of difference was either length (familiar to Korean, but not Mandarin speakers, [taja]-[tajja]) or place of articulation (retroflex or alveopalatal, familiar to Mandarin but not Korean speakers, [gotɕa]-[gotʂa]).

A Bayesian model predicts that the two groups should perform similarly on both tasks for both the similar and dissimilar words. Pajak et al. (2016)'s results were consistent with this prediction. In the discrimination task, for the highly similar words ([taja]-[tajja] or [gotɕa]-[gotʂa]), the Korean speakers outperformed the Mandarin speakers for the length contrast, and the Mandarin speakers outperformed the Korean speakers for the place contrast. However, on the word learning task, the groups did not perform differently on the two kinds of highly similar words. This suggests that although this contrast is perceptually available, something about learning both new sounds and new words simultaneously introduces a challenge.

Modeling sound and word learning as inductive inference problems highlights the necessity of input variability. Many samples are necessary in order to learn the parameters of a distribution. But having many samples is not sufficient; those samples must be representative of the actual distribution or category they are drawn from. As we have seen, acoustic variability is useful during L2 word learning in adults (Barcroft & Sommers, 2005), and some kinds of variability (but not all) are helpful for infant word learning (Rost & McMurray, 2009, 2010). What has yet to be shown is whether models of word learning within a speaker's first language benefit from similar acoustic variability.

The necessity of variability in the input makes intuitive sense to anyone who has needed to hear a word used in context many times before learning it, or who has needed years of instruction before they can tell the difference between two sounds of a new language. If all the words a learner hears in the lab are exactly the same recording repeated over many trials, this may suggest to the learner that the distribution in acoustic space is uniform, or may cue them to refrain from relying on distributional information at all.

The Switch Task

Now we will turn our attention to the Switch Task. This is an important case in which phonetic knowledge is available, and salient, but does not seem to be used during word learning. First reported by Stager and Werker (1997), this task involves teaching infants two words for two objects and then testing them on their recognition of the words. These words are typically minimal pairs, or words that differ in only one sound, such as "bih" and "dih". During training, infants are shown one object, simultaneously paired with one word, and then the second object, paired with the second word. At test, they receive a trial where the object-word pairings are the same as the training, called the *same* trial, and a trial where the object-word pairings are *different* than the training. This second kind of trial is a *switch* trial, as the words have been switched. If infants notice a change, they will spend longer looking at the switch pairing, as this is unexpected and novel.

Stager and Werker (1997) found that at 8 months, English-learning infants show discrimination between switch and same trials; when the first object is paired with the second word, they look longer at the image. At 14 months, infants do *not* show discrimination, and fail to look longer during switch trials. This is true even though they *can* hear the difference between these words in a pure discrimination task. That is, when listening to a continuous stream of "bih", if the word switches to "dih", 14-month-old infants do turn their head toward the new sound. This is surprising, because by 8 months, infants have typically tuned to the sound categories of their language, and should know that /b/ and /d/ are different sounds. Even when words differ in two features, not just one, 14-month-olds struggle with this task (Pater et al. 2004)

The difference between 8- and 14-month-olds is that by 14 months, infants have begun to learn words, that is, associate acoustic strings with objects, while this typically has not begun at 8 months. This suggests that the failure to use the available phonetic information, despite that they are *capable* of using it, is related to the word learning task. Additionally, 14-month-olds can successfully show discrimination on the same task when the words are less similar, such as "leef" and "neem" (Werker et al. 1998). This indicates that it is not just word learning, but also something about the phonetic specificity of the word representation, that contributes to their difficulty completing the switch task.

It has been proposed that this is due to task demands such as the increase in effort necessary to do both word learning and perception simultaneously (Fennell & Waxman, 2010). However, in a task with similar phonetic and word learning demands, simply changing the way learning is tested reveals that 14-month-olds can make this distinction. When tested in a 2-alternative preferential looking procedure, where there are two visual objects presented and they hear the auditory label, 14-month-olds prefer the correct object, but show a weaker preference when the label is mispronounced (Swingley & Aslin, 2002; Bailey & Plunkett, 2003)

Importantly, infants *are* making use of detailed phonetic information during this task. Rost and McMurray (2009, 2010) showed that talker variation, but not variation in phonetic features like VOT that are involved in cueing sound categories, improves infants' discrimination.

Rost and McMurray (2009) conducted the same experiment as Stager and Werker (1997), but during training, each token was produced by a different speaker. 14-month-old infants did look longer on the switch trials with this additional acoustic variability during training. However, in a follow-up study, Rost and McMurray (2010) found that multiple training tokens from the same speaker with systematically varied VOT, intended to maximize the usefulness of VOT for sound categorization, did *not* improve discrimination at test. Tokens from multiple talkers modified to each have exactly the same VOT *did* improve discrimination, indicating that not only are infants tracking talker information, they are using that information and incorporating it into their word learning process.

If infants are storing information about individual talkers, they are likely also storing category information about familiar sound categories like /b/ and /d/. It is not that infants do not know that "bih" and "dih" can be different words. Instead, perhaps the lack of talker variability is preventing appropriate generalization; infants accept mispronunciations because information from a single speaker maybe unreliable – as with speech errors or accents – even if it contains informatively variable acoustic information.

Surprisingly, when adults hear only a few tokens of a word, they – like infants – treat mispronunciations the same as correct. White et al. (2013) created a modified version of the Switch Task for adults. In this task, adults learned 48 novel words that followed English phonotactics, but were not real words of English. At test, they saw two images, one that had a trained label and one that was new, and were asked to select the image that matched what they heard. They heard either the correct label for one of the images, a mispronunciation of the correct label, or a new label. If they heard the correct label, they should select the matching image; if they heard a new label, they should select the new image. If they accept the mispronunciation as the correct label, they would select the matching image, but if they treat the mispronunciation as a new label, they would select the new image.

White et al. (2013) manipulated the number of exposures to each image-label pair during training, so the pairs were observed either 1, 5, or 8 times before test. They predicted that an image-label pair with more exposures would create a stronger lexical representation for that label and subjects would be less likely to accept a mispronunciation as correct for that image. They also ran two versions of the experiment, one where the mispronunciations differed by the correct label by a single feature of the first phoneme, and another where the mispronunciations differed by two features of the first phoneme. In the single-feature version, half of the mispronunciations were differences in manner, and half were differences in place of articulation.

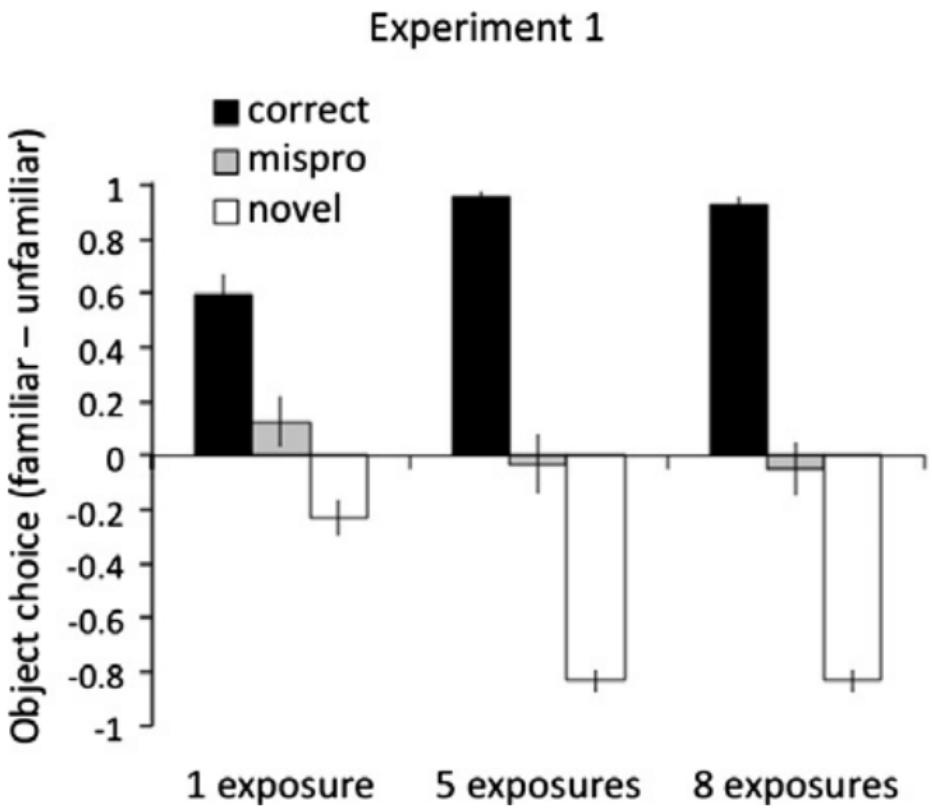

Figure 1. White et al. (2013)'s object selection results for single-feature mispronunciations (extract of Figure 2 from White et al.). Correct and Novel label selection improved with additional exposures, while Mispronounced labels did not.

As measured by proportion of selections of the correct object at test, participants differentiated between correct and mispronounced labels at all levels of frequency and mispronunciation, and this effect grew larger as frequency increased. Figure 1 shows the object selection results of their 1-feature mispronunciation experiment. In addition to participants' selections, White et al. (2013) also recorded eye-tracking data as an index of participants' sensitivity to phonetic detail. They found that for low-frequency words, participants did not show different looking behavior between correct and mispronounced labels for either 1- or 2-feature mispronunciations. Participants had different behavior between correct and mispronounced labels for higher-frequency (5 and 8 exposure) words.

This is a surprising finding; adult speakers of English know that /b/ and /d/ are different sounds, and furthermore are not in allophonic alternation in any environments of English. There is no situation in which /b/ is an acceptable variant of /d/, or vice versa. So the question is: why do English adults accept these kinds of mispronunciations?

Current Study

As White and colleagues (2013) showed, infants are not the only ones who accept inappropriate mispronunciations as correct during the early stages of word learning. Adults

similarly accept mispronunciations that are perceptually similar, but differ in ways that are not permissible alternations in their native language. In other words, they generalize their experience to a category that is too wide. When an adult English speaker learns an object is a “blem”, they often accept “glem” as an acceptable label, accepting /g/ as a possible token generated by the /b/ category, even though this does not occur anywhere else in English. The line of research presented in this paper seeks to reproduce this important finding, which has not yet been reproduced, and to investigate in what other as-yet overlooked ways adult and infant word learning might be similar.

Experiment 1 is a direct replication of the object selection task in White et al. (2013). Experiment 2 introduces single-talker variability, following Rost and McMurray (2010), where each token heard during training is a unique production by the same speaker. We hypothesize that the addition of this acoustic variability aids adults in determining the intended sound category of the speaker, and predict that this will improve adults’ rejections of mispronunciations. Note that this differs from the infant findings of Rost and McMurray (2010). We make this prediction because adults have more robust experience with the sound categories of their language, and thus the question they are faced with when hearing an unfamiliar word is different than the one infants face. For adults, the set of sound categories that a word in their language can be composed of is already determined. They additionally know which sounds are contrastive and which are allowed to vary, and along what dimensions that variability occurs. In other words, adults may need less information than infants do to firmly establish the sound categories in a word, and may benefit from acoustic variation within a single speaker.

However, if learners’ accuracy does not improve with the introduction of variability during training, it may be that it is the nature the early stages of word learning to maintain an underspecified phonetic representation until they have information from multiple sources. While this model of word learning would lead to errors of the kind we expect to find, generalizing to members outside the category, it also has benefits – an idiosyncratic speaker, an accented speaker, or even just a mistaken speaker will not completely derail the learner forever.

Experiment 1

Replicating the training-test methodology of White et al. (2013), participants learned 48 novel, English-like labels for new objects. There were two critical manipulations: number of exposures during training, and the label heard at test.

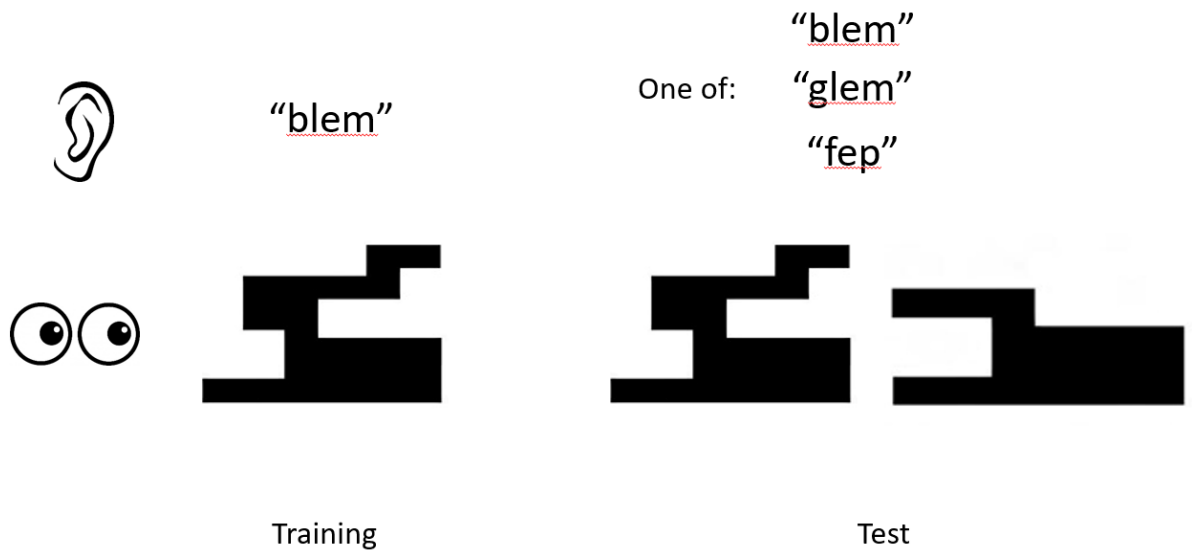

Figure 1. An illustrated example of the visual and auditory stimuli at training and test for an object-label pair.

Figure 2 illustrates a single training trial and a corresponding test trial. During the training portion, participantes saw the object on the screen and heard the label for that object, such as "blem". At test, participants saw two objects, one they had seen during training and one they had not, and heard a label. This was either the correct label, "blem", a mispronunciation, "glem", or a novel, phonetically distinct label like "fep". They were asked to click on the object that the word indicated. When the label at test was "blem", they should select the object they have seen before, which they have learned is a "blem". When they heard "fep", they should select the new object, as they have learned the other object is a "blem". The critical case is what happened when they heard "glem" – did they pick the "blem" object, indicating they accept this mispronunciation as an acceptable variant of "blem"? Or did they pick the new object, indicating "glem" is different from "blem"?

White and colleagues found that, generally, mispronunciations patterned differently from correct and novel labels – while participants selected the trained object for "blem" and the novel object for "fep" fairly accurately, they were roughly at chance on the single-feature mispronunciations, especially when they heard the word less frequently during training.

Methods

Our methods and analyses were pre-registered at on the Open Science Foundation website at https://osf.io/7vxpd/.

Participants

We conducted a power analysis using GPower's a priori power analysis for a matched pairs t-test with α =.05 and β =.8 for an effect size of F=0.5. The choice of t-test and the effect size used in this power analysis was based on the results reported in White et al. (2013)'s analysis. The effect size was estimated based on visual inspection of the difference-of-differences values in Figure 1 as the original data were not available. The estimated values were the object selection proportions for the 1-exposure (difference: 0.5) and 5-exposure (difference: 0.9) conditions, for a difference-of-difference of 0.4

Materials

Visual stimuli were identical to those used in White et al. (2013). They were geometric shapes created by generating 9x9 grids, where 18 squares in the grid were colored black and the rest white, and all colored squares bordered another colored square. There were 96 visual stimuli; 48 were paired with a label during training; the other 48 were only used during test trials. (See Figure 2 for examples of the visual stimuli.)

Auditory stimuli were single-syllable non-words following English phonotactics and using English phonemes, with either CVC or CCVC shape. The study used the same non-words as White et al. (2013). However, the audio was re-recorded by a new native speaker of English in a sound-attenuated booth and normalized to the same root-mean-squared amplitude, .075 dB SPL. (See Appendix for list of all auditory items.)

There were a total of 120 auditory stimuli: 48 used as labels during training and test, 24 only heard during test as unfamiliar test trials, and another 48 that were single-feature mispronunciations in either place of articulation or voicing (24 of each) of the trained labels. No stimuli shared the same rhyme.

Visual and auditory stimuli pairs were the same across participants. For example, all participants heard the "fep" label paired with visual presentation of Object 1, and "choom" with Object 2, etc. There were 6 stimulus lists, such that across lists, each image-label pair appeared in each exposure-by-pronunciation (correct v. mispronounced) test condition once. These image-label pairs were used as the items for by-item effects in analysis.

For the Novel test condition, participants heard novel labels and saw two images: one new image and one image they had seen, paired with a label, during training. Note that the Correct and Mispronunciation conditions fully exhausted the set of trained image-label pairs for each block. Therefore, in the Novel condition, trained images from other image-label pairs were repeated, such that Novel trials are made up of one trained and one untrained image. No two trials in the same list contained the same image-label pair (across all pronunciation test

conditions: correct, mispronunciation, and novel). The selection of trained images that were repeated for the Novel condition was counterbalanced across lists such that each trained image occurs in the Novel condition in 3 (half) of the lists.

The power analysis indicated that data from 34 participants would be necessary for an effect of our estimated size. Participants were recruited via Prolific (www.prolific.co) and paid $6 for completing the experiment. Prolific allows filtering of participants based on current geographical location; we required all participants to be current residents of the United States who had not previously completed related studies from our lab. Data were collected from 41 participants. Data from 4 participants were excluded due to self-reported cognitive impairment (n=2) or acquiring English after the age of five (n=2). None reported uncorrected speech or hearing loss. The final sample was 37 participants. Ages ranged from 19 to 49 years, with mean = 25.73. There were 17 female, 19 male, and 1 non-binary participants.

Procedure

Participants were told they would need to use a desktop or laptop computer to complete the experiment, and were asked to move to a quiet environment and use headphones. Before beginning the experiment, all participants were asked to listen to two short audio files, each with a single (real) word, and type the word they heard. They could not proceed until they correctly transcribed both words. This was to ensure that participants could, in fact, hear the auditory stimuli. After the audio test, they filled out a short questionnaire on demographics and speech, hearing, and cognitive difficulties.

The experimental paradigm was a blocked training-test procedure, where each block consisted of a training and a test portion. There were four blocks, with an opportunity to rest after each block. Participants were trained on 48 image-label pairs, 12 in each block. Breaking up training and test into four blocks reduced the difficulty of learning 48 image-label pairs, and followed the design of White et al. (2013). This is a within-subjects design; all participants were trained on all items, with the number of exposures to specific items (image-label pairs) counterbalanced across participants. The experiment itself took between twenty and thirty minutes to complete. All data was collected online in participants' browsers using a custom Javascript-based program (Chan, 2021) hosted on Google Firebase (firebase.google.com)

During each training block, each trial consisted of one image displayed in the center of the computer screen, presented simultaneously with the auditory label for that image. Training trials were self-paced with a minimum duration of 1000 ms, and participants were told to press the space bar on their keyboard to proceed to the next trial. There was a 500 ms fixation cross between each trial. Within each block, 4 of the 12 image-label pairs occurred 1 time, 4 occurred 5 times, and 4 occurred 8 times, resulting in the within-participant manipulation of exposure, for a total of 56 training trials per block.

At test, participants saw two images, one on either side of the screen, and heard a single label. One of the images was a novel image the participant had not seen before, and the other was one that received a label during training. They were instructed to press "m" on the keyboard if the label matched the image on the right side of the screen, and "z" if the label matched the image on the left side of the screen. The side of the screen each image appeared on was randomly selected for each trial. There were a total of 18 test trials per block: for each level of exposure, 2 trials had the correct label, 2 had a mispronounced label, and 2 had novel labels.

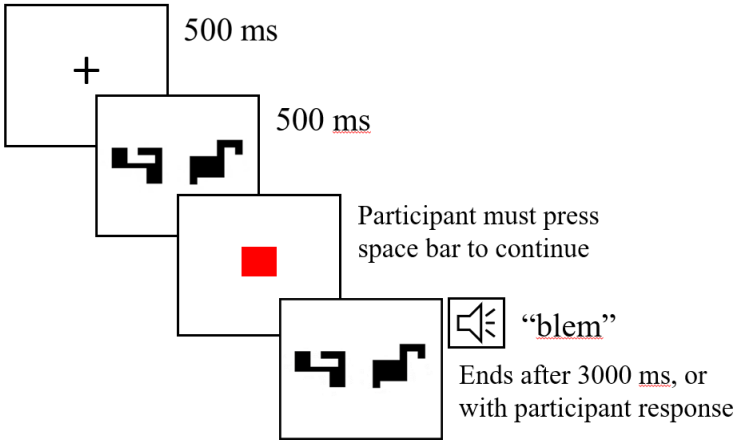

Figure 2. Test trial order of events

Test trials began with a visual preview where both images were visible for 500 ms with no audio, followed immediately by a white screen with a red box in the center. This screen would stay up until participants pressed the space bar, at which point the images would re-appear in the same locations and the auditory label was played. This visual preview was included to follow the design of White et al. (2013), who used visual preview as they were conducting both object selection and eye-tracking analyses. Participants had up to 3000 ms to select the image that matched the label they heard, and the trial terminated (i.e. moved on to the next fixation cross) after the participant made a selection. Between trials, a fixation cross appeared for 500 ms, then the next trial began. The trial sequence is visualized in Figure 3, above.

Results

Data preparation

All data preparation and analysis was conducted in R (R Core Team, 2020) and RStudio (RStudio Team, 2020). Data cleaning and analysis was performed using tidyverse (Wickham et al., 2019) packages. We began with 2610 trials, and following our pre-registered analysis plan, all trials with responses less than 100 ms or greater than or equal to 3000 ms were excluded. This constituted 132 trials, or 0.05% of the dataset, resulting in 2478 trials.

Analysis

Using the lme4 package (Bates et al., 2015), we fit a linear logistic effects model to predict accuracy with pronunciation at test (Correct vs. Mispronounced, and Mispronounced vs. Novel), exposures during training (1 vs. 5, 8 vs. 5), and all interactions between those conditions

as fixed effects. The pronunciation comparisons were treatment coded such that for the Correct vs. Mispronounced comparison, 0 = Correct, 0.5 = Mispronounced, and -0.50 = Novel. For the Mispronounced vs. Novel comparison, -0.5 = Correct, 0.5 = Mispronounced, and 0 = Novel. The exposure comparisons were similarly treatment coded, where the 1 vs. 5 comparison, 0.5 = 1 exposure, -0.5 = 5 exposures, and 0 = 8 exposures. For the 5 vs. 8 comparison, 0 = 1 exposure, -0.5 = 5 exposures, and 0.5 = 8 exposures. Results of the regression are reported in Table 1, as well as below.

We began with the maximal random effects structure, using random intercepts and slopes by participant and item (where 'item' is defined as the object-label training pair, which was consistent across all conditions and lists). The maximal random effects structure did not converge, so we removed correlation terms in order of complexity (from most to least complex, beginning with interactions). The most complex model to converge without a singular fit was one in which there were only random intercepts.

There was a significant main effect of pronunciation condition for the Mispronounced vs. Correct condition ($\beta = -1.21$, $z = -8.99$, $p < .001$), such that accuracy was lower in the Mispronounced condition. This can be seen in Figure 4, where Correct trials had higher accuracy across all three exposure conditions. This replicates a main finding of White et al. (2013), showing that participants had difficulty rejecting the mispronounced label. There was no significant overall difference in accuracy between the Mispronounced and the Novel condition ($\beta = -0.26$, $SE = .14$, $z = -1.87$, $p = 0.062$).

While there were no main effects of exposure (1 vs. 5: $\beta = -0.004$, $SE = .13$, $z = -0.034$, $p = 0.973$; 5 vs. 8: $\beta = -0.013$, $SE = .13$, $z = -0.097$, $p = .923$), exposure interacted with pronunciation condition. The difference in accuracy between the Mispronounced and Correct condition was larger at 5 exposures than at 1 exposure, ($\beta = -1.43$, $SE = .38$, $z = 3.71$, $p < .001$). In Figure 4, this can be seen in the increase in accuracy in the Correct condition between 1 and 5 exposures, while the Mispronounced condition does not see the same amount of change. Initially, additional exposures improve a learner's identification of true positives – correctly identifying the correct object for the wordform – moreso than their (correct) rejection of near phonetic neighbors. However, there was no significant change between 5 and 8 exposures ($\beta = -0.464$, $z = -1.233$, $p = .200$).

Accuracy decreased more for the Novel conditions than the Mispronounced condition between 1 and 5 exposures, $\beta = -1.43$, $z = -3.71$, $p < .001$. As seen in Figure 4, accuracy for the Novel condition decreases slightly between 1 and 5 exposures, and does so more than the Mispronounced condition. The cause of this dip in accuracy for Novel pronunciations is unclear. However, there was no significant change between 5 and 8 exposures ($\beta = 0.132$, $z = 0.38$, $p = 0.729$).

Table 1. Logistic regression table: Accuracy by pronunciation condition and number of exposures during training

Effect	Estimate	*SE*	z value	*p*
Fixed effects				
Intercept	.0543	.085	6.361	<.001*
Misp vs. Novel [a]	-.261	.140	-1.868	.062(*)
Misp vs. Corr [b]	−1.21	.135	-8.990	<.001*
1 v. 5 Exposures [c]	-.004	.130	-0.034	.973
8 v. 5 Exposures [d]	-.013	.131	-0.097	.923
Interaction Terms				
Misp vs. Novel x 1 v. 5	-1.43	.384	-3.712	<.001*
Misp vs. Novel x 8 v. 5	.132	.380	0.346	.729
Misp v. Corr x 1 v. 5 Exp	1.84	.367	5.017	<.001*
Misp v. Corr x 8 v. 5 Exp	-0.464	.376	-1.233	.217
Random effects	Variance	Std. Dev.		
Object-Label Pair	.0368	.192		
Subject	.175	.418		

Note. N = 37. CI = confidence interval; *LL* = lower limit; *UL* = upper limit. * = significant; (*) = marginal.

[a] 0 = Correct, 0.5 = Misp, -0.50 = Novel. [b] -0.5 = Correct, 0.5 = Misp, 0 = Novel. [c] 0.5 = 1 exposure, -0.5 = 5 exposures, 0 = 8 exposures. [d] 0 = 1 exposure, -0.5 = 5 exposures, 0.5 = 8 exposures.

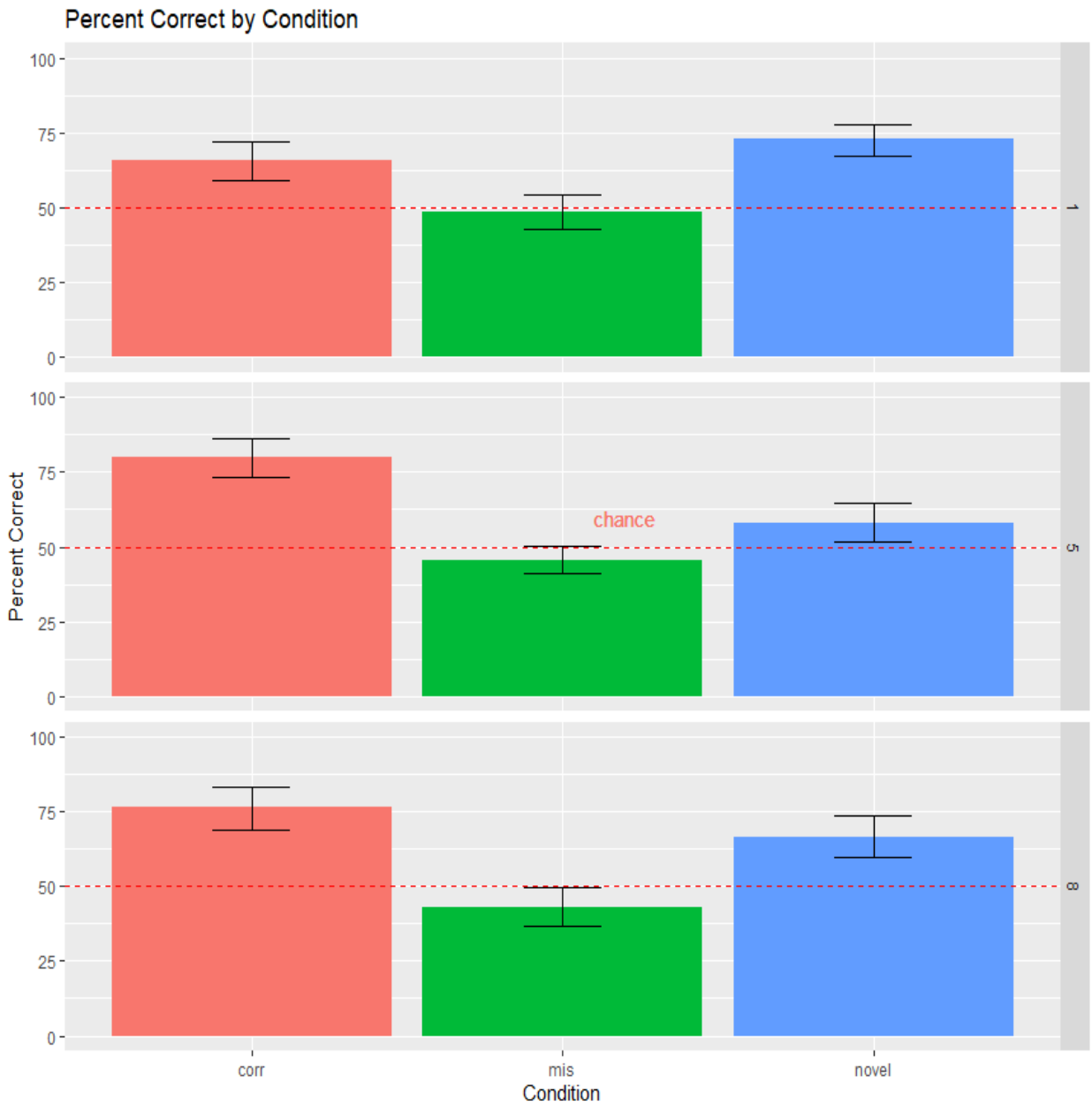

Figure 4. Percent correct by pronunciation and exposure conditions. Error bars are bootstrapped 95% confidence intervals. Chance is 50%, as each trial is a 2-alternative forced choice.

In sum, we did find a main effect of pronunciation – no matter how few exposures, participants are more accurate for the correct than the mispronounced label. This difference increases between 1 and 5 exposures, and it is driven by improvement in the correct condition, not by worse performance on mispronounced labels.

Experiment 2

Having replicated the object-selection experiment from White et al. (2013) and found a similar pattern of results, we then sought to investigate whether the addition of acoustic variability during training improved learners' correct rejections of mispronounced words at test. We ran a nearly identical experiment with the only change being in the training stimuli, which now include between 1 and 8 unique tokens of each word, spoken by the same speaker.

Methods

Participants

Participants were undergraduate students recruited through the Linguistics Department subject pool at Northwestern University and received course credit for participation (n=4) or recruited via Prolific and paid $6.00 for their participation (n=36). Data from 6 participants were excluded due to cognitive conditions (n=2), hearing loss (n=3), or acquiring English after the age of 4 (n=1). The sample before data cleaning was the target recruitment number, 34 participants. However, 3 participants did not have at least 37 usable trials after data cleaning, and were excluded for a final sample of 31. Of the final sample, 11 were female, 19 male, and 1 reported another gender. Ages ranged from 18 to 59 years, with mean = 35.42. Similar to Experiment 1, all participants were told they would need headphones, a keyboard, and a desktop or laptop computer prior to beginning the study, and completed an audio check before training began.

Materials and Procedure

All aspects of the procedure, materials, and design of this study were identical to Experiment 1, with the exception of the stimuli. The same labels from Experiment 1 were recorded 7 additional times, for a total of 8 unique tokens of each label. For the single exposure condition, the token heard by participants during training was the same token used Experiment 1. For the 5-exposure condition, four additional tokens were selected and each was heard once during training. Similarly, for the 8-exposure condition, three additional tokens were selected, and each token heard during training was unique. Tokens (aside from the Experiment 1 token) were counterbalanced, such that half of the lists used the same three tokens for the 8-exposure condition, and the other half swapped those three tokens for three from the 5-exposure condition.

Crucially, test trials in Experiment 2 were identical to test trials in Experiment 1, down to the token used for the auditory label. This means that the entire 1-exposure condition is identical between Experiment 1 and Experiment 2; participants hear the same unique token of a label in both experiments, paired with the same object, one time at training and again at test. The 5- and 8-exposure conditions differ only in that four or seven, respectively, of the tokens heard during training are new. All test trials used the Experiment 1 token.

Results

Data preparation

Data preparation and analysis followed the same procedure as Experiment 1. We began with 2360 trials and removed 239 responses outside of the response window (100-3000ms) and

all responses from participants who did not provide at least 37 usable trials, for a total of 2121 trials from Experiment 2. Data were then combined with the 2478 trials from Experiment 1, for a total of 4599 trials.

Analysis

As before, we fit a linear logistic effects regression with the maximal random effects structure by participant and item to predict accuracy. All predictors from Experiment 1 were included - pronunciation at test (Correct vs. Mispronounced, and Mispronounced vs. Novel), exposures during training (1 vs. 5, 8 vs. 5), and all interactions, coded as in Experiment 1.

Additionally, we included Experiment (and its interactions with all other predictors) as a factor, treatment coded such that Experiment 1 (single-token) = 1 and Experiment 2 (multiple-token) = 0. This will allow us to report the effects in Experiment 2 (baseline) as well as differences between Experiments 1 and 2 (interactions with this term). Thus, main effects in this regression can be interpreted as the effect of the condition within Experiment 2, the multiple-token condition. The most complex model to converge without a singular fit was, in Experiment 1, one in which there were only random intercepts.

In Experiment 2, there were two significant main effects of Pronunciation, parallel to what we observed in Experiment 1. Accuracy was lower for the Mispronounced condition than for the Novel and Correct conditions, $\beta = -0.557$, $SE = 0.152$, $z = -3.67$, $p < .001$ and $\beta = -1.136$, $SE = 0.148$, $z = -7.662$, $p < 0.001$, respectively. In contrast to Experiment 1, there was also a main effect of Exposure, with accuracy improving from 1 to 5 exposures, $\beta = -0.28$, $SE = 0.141$, $z = -1.988$, $p = 0.047$, and from 5 to 8 exposures, $\beta = 0.48$, $SE = 0.147$, $z = 3.264$, $p = 0.001$. Overall, with more training, participants improved.

Figure 5 shows the percent correct for each Pronunciation by Exposure condition, for both Experiments. The results described above can be seen in the higher accuracy for the 8-exposure condition than the 5-exposure condition in Experiment 2, as well as the comparatively poor performance on the Mispronounced condition.

Additionally, in Experiment 2 there were several significant interactions between the Pronunciation and Exposure conditions. The Mispronounced v. Novel contrast interacted with the 1 v. 5 Exposure contrast such that the difference in accuracy between the Mispronounced and Novel conditions was smaller at 1 exposure than at 5 exposures, $\beta = -1.193$, $SE = 0.415$, $z = -2.875$, $p = .004$. The same pronunciation contrast did not interact significantly with the 8 v. 5 Exposure contrast, $\beta = 0.58$, $SE = 0.424$, $z = 1.367$, $p = 0.171$. This indicates that the difference in accuracy between the Mispronounced and Novel conditions was not significantly different at 5 versus 8 exposures.

In Experiment 2, the Mispronounced v. Correct pronunciation condition significantly interacted with both the 1 v. 5 and 8 v. 5 Exposure contrasts, $\beta = 2.189$, $SE = 0.398$, $z = 5.494$, $p < 0.001$ and $\beta = -1.296$, $SE = 0.433$, $z = -2.989$, $p = 0.003$. From this, we can see that the difference between Mispronounced and Correct is smallest at 1 exposure and grows from 5 to 8

exposures. In other words, the difference between the conditions grows with exposure. Inspection of the means suggests this is driven by the correct trials.

Our main comparison of interest was between Experiment 1 and Experiment 2: did the change in stimuli during training affect participants' accuracy at test? There was no main effect of Experiment, $\beta = -0.051$, SE =0.126, $z = -0.402$, $p = 0.688$. This can be seen in Figure 5, where across all pairs of plots, the percentages are fairly similar.

However, Experiment did interact significantly with 8 v. 5 Exposure contrast, $\beta = -0.495$, SE = 0.19, $z = -2.606$, $p = 0.009$. Hearing multiple, different tokens during training increased the difference in accuracy between the 8- and 5-exposure conditions. In other words, the more tokens a participant heard, hearing a wider phonetic variety of tokens mattered more to improving accuracy.

The other two-way interaction between Exposure (1 v. 5) and Experiment was not significant, $\beta = 0.274$, SE = 0.184, $z = 1.487$, $p = 0.137$. This indicates that, in our data, having multiple tokens did not change the difference in accuracy between 1 and 5 exposures. Neither Pronunciation contrast, Mispronounced v. Novel nor Mispronounced v. Correct, interacted significantly with Experiment, $\beta = 0.297$, SE = 0.186, $z = 1.598$, $p = 0.11$ and $\beta = -0.08$, SE = 0.195, $z = -0.412$, $p = 0.68$, respectively. This can be interpreted to mean that the effect of multiple tokens during training did not have a different impact for different pronunciations at task. There were no significant three-way interactions between any of the factors (see Table 2).

Table 2. Logistic regression table: Accuracy by pronunciation condition and number of exposures during training across experiments.

Effect	Estimate	*SE*	z value	*p*
Fixed effects				
Intercept	0.595	.096	6.222	.000*
Misp v. Novel [a]	-0.557	.152	-3.666	.000*
Misp v. Corr [b]	-1.136	.148	-7.662	.000*
1 v. 5 Exposures [c]	-0.280	.141	-1.988	.047*
8 v. 5 Exposures [d]	0.480	.147	3.264	.001*
Single v. Multiple Token[e]	-0.051	.126	-0.402	.688
2-way Interaction Terms				
Misp v. Novel x 1 v. 5	-1.193	.415	-2.875	.004*
Misp v. Novel x 8 v. 5	0.580	.424	1.367	.171
Misp v. Corr x 1 v. 5	2.189	.398	5.494	.000*
Misp v. Corr x 8 v. 5	-1.296	.433	-2.989	.003*
Misp v. Novel x Single v. Multiple	0.297	.186	1.598	.110
Misp v. Corr x Single v. Multiple	-0.080	.195	-0.412	.680
1 v. 5 x Single v. Multiple	0.274	.184	1.487	.137
8 v. 5 x Single v. Multiple	-0.495	.190	-2.606	.009*
3-way Interaction Terms				
Misp v. Novel x 1 v. 5 x Single v. Multiple	-0.236	.525	-0.450	.653
Misp v. Novel x 8 v. 5 x Single v. Multiple	-0.451	.530	-0.851	.395
Misp v. Corr x 1 v. 5 x Single v. Multiple	-0.349	.532	-0.656	.512
Misp v. Corr x 8 v. 5 x Single v. Multiple	0.841	0.564	1.491	.136
Random effects	Variance	Std. Dev.		
Object-Label Pair	.037	.192		
Subject	.192	.438		

Note. N = 70. * = significant.
[a] 0 = Correct, 0.5 = Misp, -0.50 = Novel. [b] -0.5 = Correct, 0.5 = Misp, 0 = Novel. [c] 0.5 = 1 exposure, -0.5 = 5 exposures, 0 = 8 exposures. [d] 0 = 1 exposure, -0.5 = 5 exposures, 0.5 = 8 exposures. [e] 1 = Single token training (Experiment 1), 0 = Multiple token training (Experiment 2).

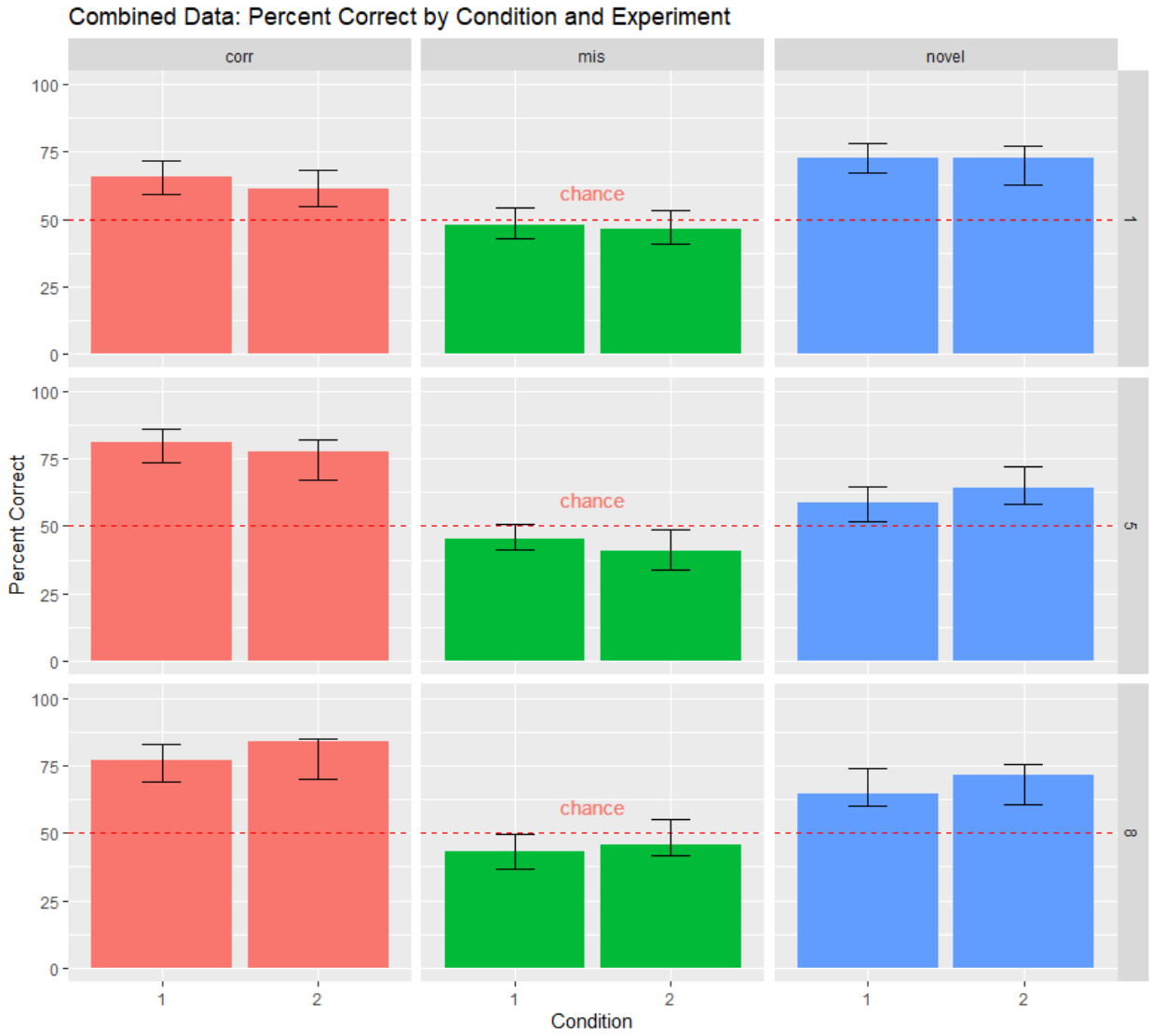

Figure 5. Percent correct by condition, exposure, and experiment. Bars on the left indicate Experiment 1, and bars on the right are Experiment 2. Chance performance is 50%, as the task was a 2-alternative forced choice.

General Discussion

In this set of experiments, we attempted to 1) replicate a previous finding that adults treat close phonetic mispronunciations similarly to correct labels during the early stages of word learning (White et al., 2013) and 2) investigate whether acoustic variability improves rejections of those same mispronunciations. We taught adult English speakers new words for novel objects, then tested whether they could correctly identify the objects that those labels referred to, varying both the number of repetitions and the label used at test within both experiments, and changing the training stimuli to multiple tokens for Experiment 2. Experiment 1 replicated prior work; learners can fairly consistently identify the correct object when they hear an identical label, or when they hear a completely dissimilar label, but perform around chance when they hear a

phonetically similar label. This suggests that word learning initially makes use of a broad, underspecified phonetic representation that encompasses not only the sounds a learner hears, but also near phonetic neighbors. The result of adding variability during training (Experiment 2) was mixed. There was no overall difference in accuracy between experiments, but when a word was heard many times (8) during training, the variability improved performance more than when the word was heard fewer times (5) during training (this showed no improvement over hearing a word just once in training).

These results are consistent with prior work. We found similar patterns of accuracy as White et al. (2013) for correct, mispronounced, and novel labels at test. Learners do well on the correct and novel labels, and struggle with the mispronounced labels, but do improve with more training across the board. The lack of overall difference between Experiment 1 and Experiment 2 also mirrors prior work with infants by Rost and McMurray (2009, 2010), which found that single-talker variation alone during the Switch Task did not improve infants' performance.

What could improve listeners' performance? Rost and McMurray (2009, 2010) found that infants benefitted from tokens from multiple talkers. Recent work by Tripp, Feldman, and Idsardi (2020, 2021) proposes that this occurs because listeners track speaker information during word learning and word recognition and, additionally, evaluate the epistemic trust of the speaker. In this instance, epistemic trust involves evaluating a speaker's reliability for a given word form. Tripp and colleagues (2020) used a Bayesian inference model to simulate the data from Rost and McMurray (2009). In their model, the acoustic realization of a wordform was generated by both the categories of the sound *and* the knowledgeability of the speaker. Their model closely mirrored the empirical results, indicating that social information, such as epistemic trust, might be driving infants' use of multi-talker variability. In follow-up simulations, the model created by Tripp and colleagues (2021) builds on a model by Shafto, Eaves, Navarro, and Perfors (2012). This model replicates Koenig and Echols (2003), which found that during a Switch Task, infants will look longer to an audio device, which was often incorrect in its labeling, than a consistently correct human speaker when each speaker provided the correct label at test. For the reliable, human speaker, a correct label is not surprising, but for the unreliable audio device, a correct label was surprising. Tripp, Feldman, and Idsardi (2021)'s model predicted looking times using both speaker reliability *and* the group the "speaker" belonged to: human or machine. The model incorporating group identity performed similarly to the infant data. This supports a model of word learning wherein speaker identity contributes to the weighting of a particular acoustic token in a word's representation. The source of the token matters, and learners know it matters.

The immediate next step in this line of research is therefore to test whether adults also benefit from multiple talkers during learning. After exploring this issue, future studies will follow up on two main questions: 1) If talker variability influences word-specific phonetic representations, what features of the talker are important? This work will specifically address the consistency of the speaker. 2) Do word learning processes change when both a word *and* a sound within the word are new to the adult learner (i.e. in the second language case)?

Conclusion

These experiments offer an initial insight into the processes involved in creating phonetic representations during word learning. Learners are willing to generalize their learning incorrectly, to accept bad examples of new words, and it is still unclear why. We are keeping track of an enormous amount of acoustic, biographical, and situational information during every speech event, when learning every new word. It's surprising that, given all that effort, we don't always use that information when it could be helpful.

References

Apfelbaum, K. S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science*, *35*(6), 1105-1138.

Ballem, K. D., & Plunkett, K. (2005). Phonological specificity in children at 1; 2. *Journal of Child Language*, *32*(1), 159-173.

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition*, *27*(3), 387-414.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1-48. doi:10.18637/jss.v067.i01.

Berk, L. E. (2003). *Child development*, *6th edition*. Allyn and Bacon.

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & psychophysics*, *61*(2), 206-219.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. I. (1997). Training Japanese listeners to identify English/r/and/l: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, *101*(4), 2299-2310.

Chan, C. (2021). Word learning experiment [Github Repository]. https://github.com/chunlchan/word-learning-experiment

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments & Computers, 28*(1), 1–11. https://doi.org/10.3758/BF03203630

Faul, F., Erdfelder, E., Lang, AG. *et al.* G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* **39,** 175–191 (2007). https://doi.org/10.3758/BF03193146

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological review*, *120*(4), 751.

Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., & Morgan, J. L. (2013). Word-level information influences phonetic learning in adults and infants. *Cognition*, *127*(3), 427-438.

Fennell, C. T., & Waxman, S. R. (2010). What paradox? Referential cues allow for infant use of phonetic detail in word learning. *Child development*, *81*(5), 1376-1383.

Foulkes, P., & Hay, J. B. (2015). 13 The Emergence of Sociophonetic Structure. *The Handbook of Language Emergence*, *87*, 292.

Ganger, J., & Brent, M. R. (2004). Reexamining the vocabulary spurt. *Developmental psychology*, *40*(4), 621.

Goldinger, S. D., & Azuma, T. (2004). Episodic memory reflected in printed word naming. *Psychonomic bulletin & review*, *11*(4), 716-722.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical society of America*, *97*(5), 3099-3111.

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help?. *Language, cognition and neuroscience*, *34*(1), 43-68.

Koenig, M. A., & Echols, C. H. (2003). Infants' understanding of false labeling events: The referential roles of words and the speakers who use them. *Cognition*, *87*(3), 179-208.

Kovack-Lesh, K. A., & Oakes, L. M. (2007). Hold your horses: How exposure to different items influences infant categorization. *Journal of Experimental Child Psychology*, *98*(2), 69-93.

Lev-Ari, S. (2017). Talking to fewer people leads to having more malleable linguistic representations. *PloS one*, *12*(8), e0183593.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), B33-B42.

Milton, J., & Treffers-Daller, J. (2013). Vocabulary size revisited: the link between vocabulary size and academic achievement. *Applied Linguistics Review*, *4*(1), 151-172.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The journal of the acoustical society of America*, *85*(1), 365-378.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & psychophysics*, *60*(3), 355-376.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Perception & psychophysics*, *57*(7), 989-1001.

Pajak, B., Creel, S. C., & Levy, R. (2016). Difficulty in learning similar-sounding words: A developmental stage or a general property of learning?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(9), 1377.

Pajak, B., Fine, A. B., Kleinschmidt, D. F., & Jaeger, T. F. (2016). Learning additional languages as hierarchical probabilistic inference: Insights from first language processing. *Language Learning*, *66*(4), 900-944.

Palan, S. & Schitter, C. (2018). Prolific.ac - A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17* (2018), pp. 22-27. doi:10.1016/j.jbef.2017.12.004

Prolific.co. (2021). *Prolific | Online participant recruitment for surveys and market research* [online]. Available at: https://www.prolific.co/

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental science*, *12*(2), 339-349.

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy*, *15*(6), 608-635.

RStudio Team (2020). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental science*, *15*(3), 436-447.

Singh, L., White, K. S., & Morgan, J. L. (2008). Building a word-form lexicon in the face of variable input: Influences of pitch and amplitude on early spoken word recognition. *Language Learning and Development*, *4*(2), 157-178.

Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1994). Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude. *The Journal of the Acoustical Society of America*, *96*(3), 1314-1324.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*(6640), 381-382.

Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological science*, *13*(5), 480-484.

Thompson, B., & De Boer, B. (2017). Structure and abstraction in phonetic computation: Learning to generalise among concurrent acquisition problems. *Journal of Language Evolution*, *2*(1), 94-112.

Tingsabadh, M. K., & Abramson, A. S. (1993). Thai. *Journal of the International Phonetic Association*, *23*(1), 24-28.

Tripp, A., Feldman, N. H., & Idsardi, W. J. (2021). Social Inference May Guide Early Lexical Learning. *Frontiers in Psychology*, *12*.

Tripp, A., Feldman, N., & Idsardi, W. (2020, April). Are infants sensitive to informant reliability in word learning?. In *Proceedings of the Annual Boston University Conference on Language Development*.

Twomey, K. E., Ma, L., & Westermann, G. (2018). All the right noises: Background variability helps early word learning. *Cognitive Science*, *42*, 413-438.

Twomey, K. E., Ranson, S. L., & Horst, J. S. (2014). That's more like it: Multiple exemplars facilitate word learning. *Infant and Child Development*, *23*(2), 105-122.

Verma J.P., Verma P. (2020) Use of G*Power Software. In: Determining Sample Size and Power in Research Studies. Springer, Singapore. https://doi.org/10.1007/978-981-15-5204-5_5

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development*, *7*(1), 49-63.

White, K. S., Yee, E., Blumstein, S. E., & Morgan, J. L. (2013). Adults show less sensitivity to phonetic detail in unfamiliar words, too. *Journal of Memory and Language*, *68*(4), 362-378.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, doi:10.21105/joss.01686

Appendix

Table A1. International Phonetic Alphabet pronunciations of the words heard by participants in both experiments.

Correct Label	Mispronounced Label
bæv	gæv
bith	gith
blɛm	glɛm
boɪð	goɪð
bos	pos
buʒ	puʒ
baʊdʒ	paʊdʒ
daʃ	taʃ
dɛtʃ	bɛtʃ
ðɛk	ʒɛk
dɪv	bɪv
dreb	greb
fɪdʒ	vɪdʒ
faʊm	vaʊm
gədʒ	kədʒ
grol	brol
gaɪn	baɪn
gef	kef
kadʒ	padʒ
kɛl	gɛl
kɪʒ	gɪʒ
gʊz	dʊz
klop	glop
pəv	bəv
pluk	kluk
pʊth	tʊth
sɛp	fɛp
sig	ʃig
soɪk	zoɪk
sot	zot
ʃub	ʒub
sʊd	zʊd
saɪp	ʃaɪp
tæs	kæs
θəp	fəp
teg	peg

tib	dib
tof	dof
toɪp	doɪp
tram	dram
tuv	kuv
vad	zad
voɪn	Zoɪn
vup	zup
zəl	səl
zæb	sæb
zed	sed
ʒib	ʃib

Novel Labels
tʃæg
flɪm
slud
dʒɪk
ʃɛn
kaʊt
faɪs
san
vɛd
taɪf
vis
daɪg
præk
stət
læt
gaʊk
tɛz
tʃum
tʃəb
paɪtʃ
dʒid
fæl
voʒ
fitʃ