Acquiring Phonemes: Is Frequency or the Lexicon the Dominant Cue?

Emily Moeng
moeng@live.unc.edu

Acquiring Phonemes: Is Frequency or the Lexicon the Dominant Cue?

Acquiring Phonemes: Is Frequency or the Lexicon the Dominant Cue?

Emily Moeng

1 Introduction

One of the most influential proposals concerning how phonemes are acquired claims that language learners track frequencies of sound tokens to infer the number of phonemes in the ambient language (Maye et al. 2002). Although this **frequency-based account** has been influential among acquisitionists (e.g., Gervain and Mehler 2010), it has been claimed that this account alone is unable to arrive at the correct number of phonemes when given data taken from natural language, especially when it comes to vowels. This has prompted some researchers to suggest a **lexicon-based account**, in which infants utilize high-frequency lexical items to aid the acquisition of phonemes. While both the frequency- and lexicon-based accounts have been experimentally supported, no study has yet compared the interaction of these two effects. This study is concerned with determining which of these two is used as the dominant cue by language learners. More specifically, this study's research questions are: *When the Lexical and Frequency Cue give the learner conflicting information…* (1) Which is treated as dominant? (2) Do learners rely on different cues when determining vowel and stop categories? (3) Do learners rely on one cue early on and the other later?

This study presents adults with an artificial language in which the Lexical Cue and the Frequency Cue give the learner conflicting information regarding the number of phonemes. It is found that the Lexical Cue has a non-significant tendency to be treated as the dominant cue, but only for the vowel stimuli. Consonants on the other hand showed no consistent trend.

2 Background

Phonemes are largely arbitrary and must be acquired. Infants exhibit language-specific discrimination of consonants around 10 months (Werker and Tees 1984), and of vowels around 6 months (Kuhl et al. 1992). Introduced in this section are the two main proposals for how phonemes are acquired so early: the Frequency Cue, and the Lexical Cue.

2.1 The Frequency Cue ("Distributional Learning")

One account for how learners acquire phonemes is known as the **distributional learning** hypothesis, referred to here as the **frequency-based learning** hypothesis[1]. According to this theory, learners map tokens into some phonetic space, and use their relative frequencies to infer the number of phonemes in the ambient language. A learner exposed to a bimodal distribution will infer that there are two phonemes; a learner exposed to a monomodal distribution will infer that there is one.

Studies show that learners are capable of the computations necessary to utilize this proposed Frequency Cue. Maye and Gerken (2000) found that a group of participants exposed to an artificial language with a bimodal distribution of tokens ranging between a voiceless unaspirated stop [t] (like in *s**t**eam*, not ***t**eam*) and a pre-voiced stop [d] (like in ***d**eem*) inferred that there were two categories, whereas a group exposed to a monomodal distribution inferred that there was only one.

Therefore, learners receive what will be referred to here as a **Frequency Cue** informing them of the number of phonemes in the ambient language. If a learner is exposed to a MONOMODAL distribution, (s)he will receive a Frequency Cue that there is a single phoneme. If a learner is exposed to a BIMODAL distribution, (s)he will receive a Frequency Cue that there are two phonemes.

Maye et al.'s findings, although widely cited (e.g., there are 663 Google Scholar citations of Maye et al. (2002), as of this writing), have been replicated, but not with an extensive variety of test stimuli. Experimental support has been found for adults (e.g., Maye and Gerken 2000) and infants (Maye et al. 2002). Participants are able to generalize from synthetic to natural speech (Gulian et al., 2007), but not the ability to generalize using features (Maye and Gerken 2001).

[1] Though usually referred to as "distributional learning," I will be referring to it as the frequency-based learning, since I later also refer to environmental, rather than frequency-related, distribution.

Attempts to replicate Maye and Gerken's (2000) findings to other stimuli have shown mixed success. Stimuli successfully used in replications have included consonants ranging from a prevoiced stop [d] to a voiceless unaspirated stop [t] (Maye and Gerken 2000, Maye and Gerken 2001, Maye et al. 2002), consonants ranging from prevoiced [g] to voiceless unaspirated stop [k] (Hayes-Harb 2007), vowels ranging from [a] to [ɑ] (Gulian et al. 2007), and vowels ranging from [i] to [ɪ] (Gulian et al. 2007). However, Peperkamp et al. (2003) failed to replicate these findings when using the fricatives [ʁ] – [χ] with French-speaking adults.

2.2 Why the Frequency Cue Alone is Not Sufficient: Vowel Categories Overlap

These artificial language learning tasks have shown that both infants and adults are able to make the required statistical calculations when given simplified data. The question now becomes whether natural language exhibits a distribution of phonemes similar to the distribution of phonemes in these artificial languages. Unfortunately for a purely frequency-based distributional account, this does not seem to be the case. As demonstrated by Swingley (2009) in Figure 1, vowels overlap to such an extent that the clear clusters predicted by a frequency-based account are not visible.

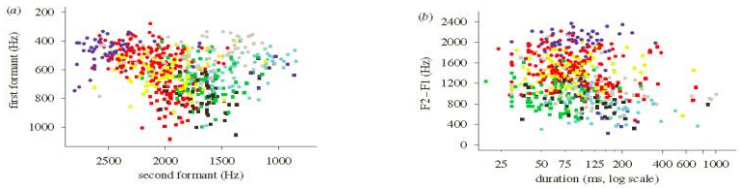

Figure 1. Plot of 11 English monophthongs from Swingley (2009) demonstrating the "overlapping categories" problem for a frequency account, using different phonetic spaces: (a) F1 vs. F2 and (b) F2-F1 vs. duration. According to a purely frequency-based account, 11 clusters should be visible.

Therefore we can sum up the main problem with a purely frequency-based account as a problem of **overlapping categories** in natural linguistic conditions. (See also Bion et al. 2013.) However, this issue of overlapping categories might only apply to some phonemes. As seen in Figure 2a, peaks do form if we look at voice onset time (VOT) of stops. English speakers are exposed to a distribution with two prominent peaks and one smaller peak. We could imagine either a model in which learners only notice more prominent peaks in frequency (i.e., frequencies that fall above some threshold), or a model in which learners notice all local maxima in frequency (three in this case), and then, through some second step, collapse categories which are in complementary distribution into a single category. By comparison, Dutch speakers (Figure 2b), who have a single phoneme associated with velar stops, are exposed to only a single peak in frequency.

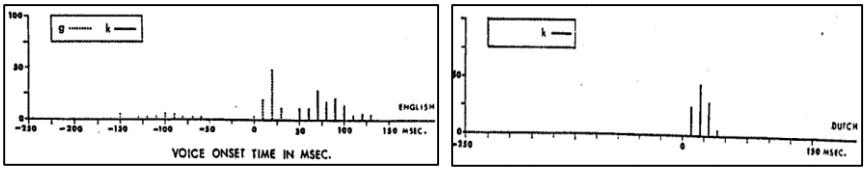

Figure 2a and b. Lisker and Abramson (1964). In a figure plotting VOT in [k] and [g], we see (a) two or three clear peaks form in English, and (b) only one peak form in Dutch.

2.3 The Lexical Cue

One possible solution to the "overlapping categories" problem is that learners do not acquire phonemes in isolation from learning words. Rather than acquiring phonemes *before* learning words, learners may use frequently-heard words to aid them in phoneme acquisition. This will be referred to as **lexicon-based learning**.

A lexicon-based account claims that learners retain the lexical origins of individual phonemes, and use these as guides when creating phoneme boundaries. At this early point in acquisition, in-

fants begin with the assumption that a pair of words are different lexical items only if their acoustic signals overall are very different from one another. That is, infants begin with a bias *against* the existence of minimal pairs. For example, [da] and [tagu] are likely different "words," but by this assumption, an infant would assume that [da] and [ta] are the same "word," even if shown to have different semantic references. This is supported by Stager and Werker's (1997) finding that 14-month olds habituated on object-label pairings for *bih* and for *dih* failed to notice when the object-label pairings were switched, despite being able to discriminate between [b] and [d].

Thiessen (2007) and Feldman et al. (2011, 2013) provide experimental support that infants and adults make use of word-level information when acquiring phonemes. Thiessen habituated 15-month old infants on 3 sound-meaning pairs: a test word *daw*, and two dissimilar-sounding phonetic forms (*tawgoo*, and *dawbow*)[2]. Each word was paired with a single visual object. Another group was habituated on 3 sound-meaning pairs: a test word *daw*, and two **similar**-sounding phonetic forms (*tawgoo* and *dawgoo*). Again, each was paired with a single visual object. Thiessen found that infants trained on words in which [d] and [t] occurred in different lexical contexts (*tawgoo* and *dawbow*) were more likely to notice a switch in sound-object pairings of *daw* to *taw*, than infants trained on words in which [d] and [t] occurred in the same contexts (*tawgoo* and *dawgoo*).

Feldman et al. (2011) familiarized adults on bisyllabic words ending in a syllable taken from an 8-point continuum ranging between [tɑ] and [tɔ]. tA_1 refers to the most [tɑ]-like end of the continuum, while tA_8 refers to the most [tɔ]-like end. 2 groups heard words from an artificial language. One group, referred to here as the NOSAMEENVIRONMENTS group, heard the words $gutA_{1\text{-}4}$ and $litA_{5\text{-}8}$, or the words $gutA_{5\text{-}8}$ and $litA_{1\text{-}4}$, as well as a number of filler words. Therefore this group never heard $tA_{1\text{-}4}$ and $tA_{5\text{-}8}$ in the same lexical environment. The other group, referred to here as the SAMEENVIRONMENTS group, heard the words $gutA_{1\text{-}4}$, $gutA_{5\text{-}8}$, $litA_{1\text{-}4}$, and $litA_{5\text{-}8}$, as well as a number of fillers. Therefore this group heard $tA_{1\text{-}4}$ and $tA_{5\text{-}8}$ in the same lexical environments. The authors found that the NOSAMEENVIRONMENTS group was more likely to respond that [tɑ] and [tɔ] were "different" from one another, as compared to the SAMEENVIRONMENTS group. This result was found for adults (Feldman et al. 2011), as well as for 8-month olds (Feldman et al. 2013).

Therefore, in addition to receiving a Frequency Cue, language learners also receive what will be referred to here as a **Lexical Cue** informing them of the number of phonemes in the ambient language. If a learner is exposed to NOSAMEENVIRONMENTS lexical items, (s)he will receive a Lexical Cue that there is a single phoneme. If a learner is exposed to a SAMEENVIRONMENTS lexical items, (s)he will receive a Lexical Cue that there are two phonemes.

3 Research question

This study's main questions are: *When the Lexical and Frequency Cue give the learner conflicting information…* (1) Which is treated as dominant? (2) Do learners rely on different cues when determining vowel and stop categories? (3) Do learners rely on one cue early on and the other later?

4 Design

The experiment design followed the design of Maye and Gerken (2000), who provided experimental support for the Frequency Cue, as well as the design of Feldman et al. (2013), who provided experimental support for the Lexical Cue. Although the experiments for both Maye and Gerken (2000) and Feldman et al. (2013) lasted for only a single session, this experiment tested participants over three days, spaced at least 18 hours apart. All three days consisted of identical procedures: a Familiarization phase followed by a Test phase. There were four independent variables:

(1) **Consonant vs. Vowel *(within-subject)***: This experiment tested consonants (drawn from an 8-point continuum between [t] – [t^h]) and vowels (drawn from [ɑ] – [ɔ]).
(2) **Lexical Cue *(across-subject)***: Like in Feldman et al. (2013), learners were exposed to test tokens embedded within two-syllable lexical contexts. The SAMEENVIRONMENTS group heard test tokens embedded within the same lexical context ($limsA_{1\text{-}4}$ and $limsA_{5\text{-}8}$), whereas

[2] Orthography from Thiessen (2007). Exact IPA transcriptions are unknown.

the NOSAMEENVIRONMENTS group ***never*** heard test tokens embedded within the same lexical context (e.g., *limsA*$_{1\text{-}4}$, but never *limsA*$_{5\text{-}8}$).

(3) **Frequency Cue *(across-subject)***: Like in Maye and Gerken (2000), learners were exposed to different frequencies of test tokens: BIMODAL or MONOMODAL.

(4) **Early vs. Late *(within-subject)***: This experiment consisted of three identical sessions (a Familiarization phase followed by a Test phase), spaced at least 18 hours apart.

Test words for each condition are shown in Figure 3. Within each cell, the *x*-axis indicates a phonetic continuum, either between [t] and [t^{h}], or between [ɑ] and [ɔ]. The *y*-axis indicates each token's frequency: one peak indicates a monomodal distribution; two peaks indicate a bimodal distribution. The two groups tested in this experiment were the (i) BIMODAL SAMEENVIRONMENTS, and (ii) MONOMODAL NOSAMEENVIRONMENTS groups, the only two cells in which the Lexical Cue and the Frequency Cue give the language learner conflicting information.

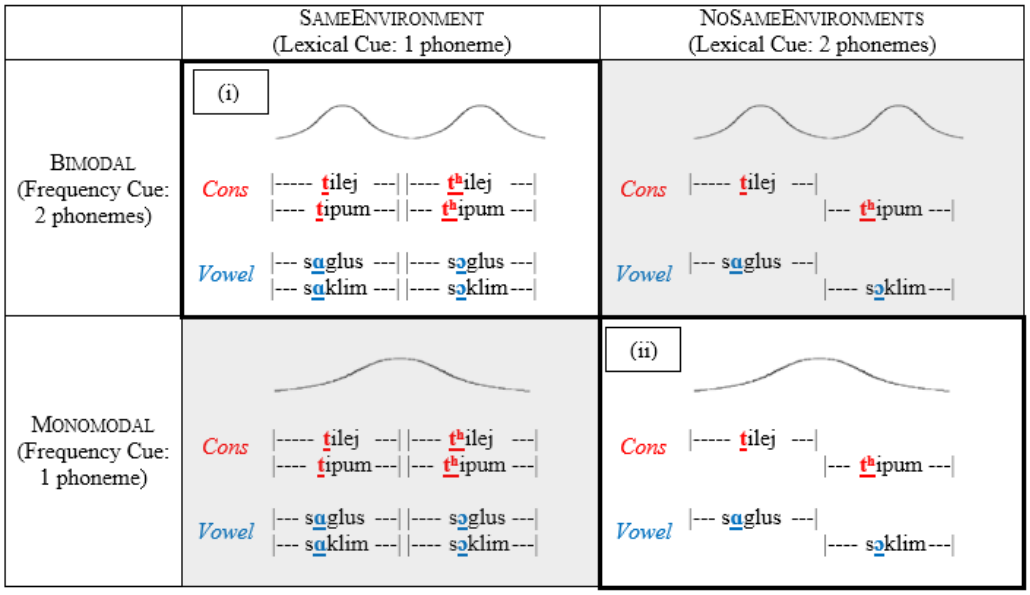

Figure 3. Experiment design. The two groups tested in the present study were cell (i) BIMODAL SAMEENVIRONMENTS and cell (ii) MONOMODAL NOSAMEENVIRONMENTS.

5 Method

5.1 Participants

Participants were recruited from Mechanical Turk, an online participant pool (see Crump et al. (2013) for a discussion on the legitimacy of conducting psychological experiments through MTurk). Participants were located in the United States, and were asked to participate only if they: (1) had no known history of a speech/hearing impairment, (2) were 18 or older, (3) were native English speakers, (4) had regular access to the internet, (5) could play audio on their computer.

5.2 Stimuli

There were three types of stimuli used in the Familiarization phase of this experiment: filler words, test consonant words, and test vowel words. There were two types of stimuli used in the Test phase of this experiment: control syllables and test syllables. Stimuli were recorded by the experimenter, a native speaker of English. Recordings were made in a soundproof booth on an Acer netbook at 44100 Hz, using Praat, a piece of speech analysis software (Boersma and Weenink, 2013).

5.2.1 Familiarization Phase: Filler Words

Filler words were bisyllabic pseudowords, recorded by the experimenter, a native English speaker.

5.2.2 Familiarization Phase: Test Consonants

Test consonants were drawn from an 8-point continuum between [t] and [t^h] (*T₁i – T₈i* and *T₁u – T₈u*). The *T₁u –T₈u* continuum was constructed by removing the aspiration from the end of a [t^hu] token (which originally had 65 ms of aspiration). The continuum was created so that the end point *T₁* had the amount of aspiration found in a naturally produced [stu] syllable (11 ms), and so that, as judged by the experimenter (a native English speaker), it sounded like "too" was switching to "doo" around the *T₄/T₅* midpoint of the continuum. The *T₁i – T₈i* continuum was created in the same way. Specific amounts of aspiration for each point are listed in Table 1a.

Point along continuum	Amount of aspiration (ms)
T_1 (t)	11
T_2	18
T_3	25
T_4	32
T_5	39
T_6	46
T_7	53
T_8 (t^h)	60

Point along continuum	F2 (Hz)
A_1 (ɑ)	1278
A_2	1220
A_3	1163
A_4	1105
A_5	1047
A_6	989
A_7	931
A_8 (ɔ)	873

Table 1. (a) Amount of aspiration for each point along the T_1-T_8 continuum. (b) F2 for each point along the A_1-A_8 continuum.

These *Ti* and *Tu* syllables were 1) added before a contextual syllable spliced out of a naturally-produced word of the form [t^hə.***(syll)***], and 2) added after a contextual syllable spliced from a naturally-produced word of the form [***(syll)***.t^hə]. Cuts were made at zero crossings to avoid clicks. This created 16 *T*-words, 4 of each of the following types: (syll)Ti, (syll)Tu, Ti(syll), and Tu(syll).

5.2.3 Familiarization Phase: Test Vowels

Test vowels were drawn from an 8-point continuum between [ɑ] and [ɔ] (s*A₁ – sA₈* and z*A₁ – zA₈*). The *A₁ – A₈* continuum was made by manipulating formants of a naturally-produced [ɔ]. This was done in Praat by editing the oral formant grid in Praat's simple KlattGrid. KlattGrid is a speech synthesizer built into Praat (see Weenink 2009). The only parameters manipulated in this experiment were the first 5 oral formant values.

Formants at the very beginning of the vowel were unaltered, and formant values for the rest of the vowel (starting approximately 10% of the way into the vowel) were manipulated so that they had the following steady values: F1=800, F3=2800, F4=4000, F5=4500. Only F2 values differed (Table 1b). F2 values were chosen so that the *A₁* endpoint sounded like an [ɑ] (as judged by the experimenter), and so that it sounded as if "ah" was switching to "aw" around the *A₄/A₅* midpoint of the continuum. Each vowel in this 8-point continuum was then spliced to an [s] removed from [sɔ] and a [z] removed from [zɔ] to produce *sA* and *zA* syllables. These *sA* and *zA* syllables were then 1) added before a contextual syllable from a naturally-produced word of the form [sə.***(syll)***], and 2) added after a contextual syllable from a naturally-produced word of the form [***(syll)***.sə]. This produced 16 *A*-words, 4 of each of the following the types: (syll)sA; (syll)zA; sA(syll); and zA(syll).

To summarize, 16 two-syllable *T*-words and 16 two-syllable *A*-words were created, where each "word" had 8 variations corresponding to the 8 continuum points (T_{1-8} or A_{1-8}). After creating all words, each was filtered in Praat using a pre-emphasis filter from a frequency of 50 Hz.[3] This increased the spectral slope by 6 dB/octave above the frequency of 50 Hz.

[3] The resynthesized vowel sounded robotic and unnatural compared to the unaltered speech, so a filter was applied to make the entire word sound more robotic. However, it should be noted that, even with the filter applied, the resynthesized vowels did stand out a bit from the unaltered speech.

5.2.4 Test Phase: Control Syllables and Test Syllables

The control syllables for the Test phase consisted of one-syllable CV words recorded by the experimenter. "Same" control syllables were two different recordings of the same syllable (ex: $[gi]_1$ and $[gi]_2$). "Different" control syllables were recordings of two different syllables (ex: [gi] and [bi]).

Test syllables for the Test phase consisted of a *zA* continuum, and a *Tu* continuum, each created from different recordings from the previous *zA* and *Tu* continua, but by using the same methods as described in Section 5.2.2 and 5.2.3.

5.3 Procedure

Each day consisted of one Familiarization phase, followed by one Test phase. At the end of the third day (i.e. the last day), participants were asked to answer a short questionnaire.

5.3.1 Familiarization Phase

In the Familiarization phase, participants heard 96 experimental tokens and 33 fillers. Figure4a shows example stimuli that a participant in the BIMODAL SAMEENVIRONMENTS group could have heard during the Familiarization phase. The specific tokens heard varied randomly for each participant, but all followed the following two constraints: 1) each test word was heard three times, and 2) each point along the 8-point continuum was heard such that a bimodal distribution was created.

A participant in the MONOMODAL NOSAMEENVIRONMENTS group also heard 96 experimental tokens and 33 filler tokens in the Familiarization phase. Figure4b shows example stimuli that a participant in the MONOMODAL NOSAMEENVIRONMENTS group could have heard during the Familiarization phase. Specific tokens heard varied for each participant, but all followed three constraints: 1) each test word was heard three times, 2) points 1-4 in the *T*- or *A*-continua were never heard in the same lexical environment as the points 5-8 in the same *T*- or *A*-continua, and 3) each point along the 8-point continuum was heard such that a ***mono***modal distribution was created.

To ensure participants were paying attention during the Familiarization phase, filler words were followed by a bell sound, while non-filler words were not followed by a bell sound. Participants were asked to click either "word and bell" or "word only" to indicate whether they had heard a bell sound. If participants answered incorrectly more than 6 times (i.e. if they answered "word and bell" when the sound of a bell had not played, or "word only" when the sound of a bell had played), their results were excluded from further analysis.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	Total
limsA		2			1				3
mæsA			1			1	1		3
nejʤsA		1		1		1			3
spilsA		1					2		3
nulzA		1			1		1		3
pibzA			1				1	1	3
rejmzA			1			1	1		3
sæmzA	1	1		1					3
Total (Bimodal)	1	6	3	2	2	3	6	1	

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	Total
limsA	1			2					3
mæsA		1	1	1					3
nejʤsA					2	1			3
spilsA					1	1	1		3
nulzA		1		2					3
pibzA			2	1					3
rejmzA					1	1	1		3
sæmzA					2			1	3
Total (Monomodal)	1	2	3	6	6	3	2	1	

Figure 4. (a) Sample Familiarization for *(syll)sA* and *(syll)zA* stimuli for a participant in the BIMODAL SAMEENVIRONMENTS group (Frequency Cue => 2 categories; Lexical Cue => 1 category). (b) Sample Familiarization for *(syll)sA* and *(syll)zA* stimuli for a MONOMODAL NOSAMEENVIRONMENTS participant (Frequency Cue => 1 category; Lexical Cue => 2 categories).

5.3.2 Test Phase

After the Familiarization phase participants were directed to a Test phase. Here they were then given pairs of syllables and asked if they were the same or different[4]. They heard 2 repetitions of

[4] Following Feldman et al. (2009), participants were told they should answer "different" if the pairs of syllables had different sound like in English CAP and GAP, and answer "same" if the pairs of syllables were

the following types of pairs: 8 Control Pairs[5] (e.g., *ni* vs. *nu*, *ni* vs. *ni…*); Far Contrast (continuum point 2 vs. continuum point 7) (i.e., T_2u vs. T_7u and zA_2 vs. zA_7); Near Contrast (point 3 vs. point 6) (i.e., T_3u vs. T_6u and zA_3 vs. zA_6); and Within-Category Contrast (point 1 vs. point 4, and point 5 vs. point 8) (i.e., T_1u vs. T_4u, T_5u vs. T_8u, zA_1 vs. zA_4, and zA_5 vs. zA_8).

5.3.3 Questionnaire (End of Day 3)

18 hours after completing Day 1, participants were eligible to participate in the identical Day 2 experiment. 18 hours after completing Day 2, they were eligible for Day 3. At the end of Day 3, participants were asked (among other things) whether they pronounced the words *cot* and *caught* the same. People who are said to "have" the low back vowel merger pronounce /ɑ/ and /ɔ/ the same, resulting in the homophones *cot* and *caught* (e.g., Vaux and Golder, 2003).

6 Results

143 people participated in this experiment. Of those, 27 were excluded from the results, either due to a technical difficulty (7 participants) or because they did not pass tests designed to see whether they were paying attention (20 participants)[6], leaving a total of 116 participants. A total of 61 participants participated for all three days. A population average logistic regression model was fitted with the GENMOD Procedure in SAS 9.3, accounting for multiple observations within subjects.

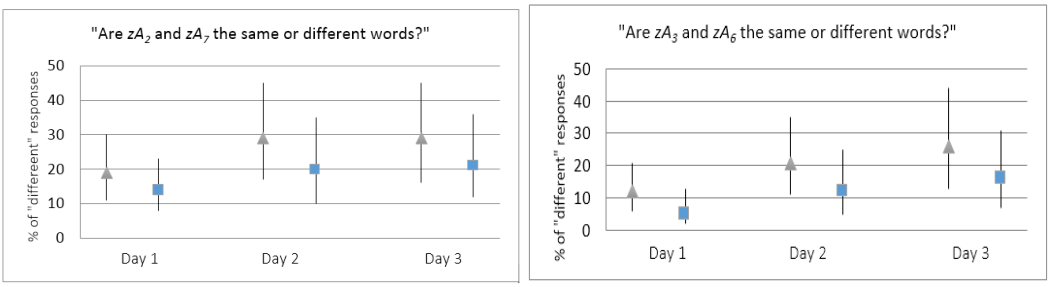

Figure 5. Vowel results for the MONOMODAL NOSAMEENVIRONMENTS (grey triangles) and BIMODAL SAMEENVIRONMENTS (blue squares) groups. (a) Far contrast, A_2 vs. A_7. (b) Near contrast, A_3 vs. A_6.

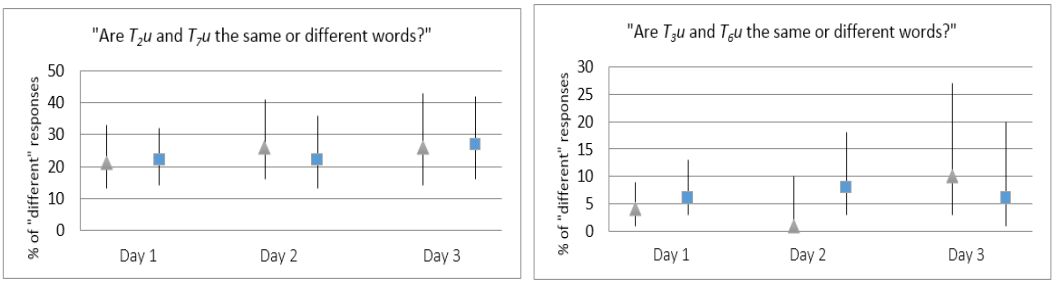

Figure 6. Consonant results for the MONOMODAL NOSAMEENVIRONMENTS (grey triangles) and BIMODAL SAMEENVIRONMENTS (blue squares) groups. (a) Far contrast, T_2 vs. T_7. (b) Near contrast, T_3 vs. T_6.

---

the same, even if pronounced slightly differently, like two pronunciations of GAP and GAP.

[5] The "same" control consisted of 2 different recordings of the same syllable (e.g., 2 [nu] recordings).

[6] Participants were excluded 1) if they answered incorrectly more than 6 (out of 129) times in the Training phase *(8 participants were excluded from analysis due to this criterion)*, or 2) if they answered "same" more than 2 (out of 8) times in the Test phase for the control words that were different (*ni* vs. *nu*) *(4 participants were excluded from analysis due to this criterion)*, or answered "different" more than 2 (out of 8) times in the Test phase for the control words that were the same (*ni* vs. *ni*) *(8 participants were excluded from analysis due to this criterion)*.

Figure 5 shows the results of the far (A_2 vs. A_7) and near (A_3 vs. A_6) vowel contrast. Error bars represent a 95% confidence interval. Although not significantly differing from one another (p=0.1876), it can be seen that, consistently across all three days and for both contrasts tested, the MONOMODAL NOSAMEENVIRONMENTS group (grey triangles) answers "different" more often than the BIMODAL SAMEENVIRONMENTS group (blue squares).

Figure 6shows the results of the far (T_2 vs. T_7) and near (T_3 vs. T_6) consonant contrast. Error bars represent a 95% confidence interval. Unlike the vowel results, no consistent trend is observed.

6.1 A Note on the Cot-Caught Merger

It is believed that the cot-caught merger will have an effect on participant responses, since participants with the merger (i.e., those who pronounce /ɑ/ and /ɔ/ the same) are being taught a contrast, whereas those without the merger (i.e., those who pronounce /ɑ/ and /ɔ/ differently) are being taught to ignore a contrast that they have. Therefore it could be the case that the tendency for the MONOMODAL NOSAMEENVIRONMENTS group to answer "different" more often than the BIMODAL SAMEENVIRONMENTS group can be attributed to a greater proportion of people with the *cot-caught* merger in the BIMODAL SAMEENVIRONMENTS group. This section tests that hypothesis.

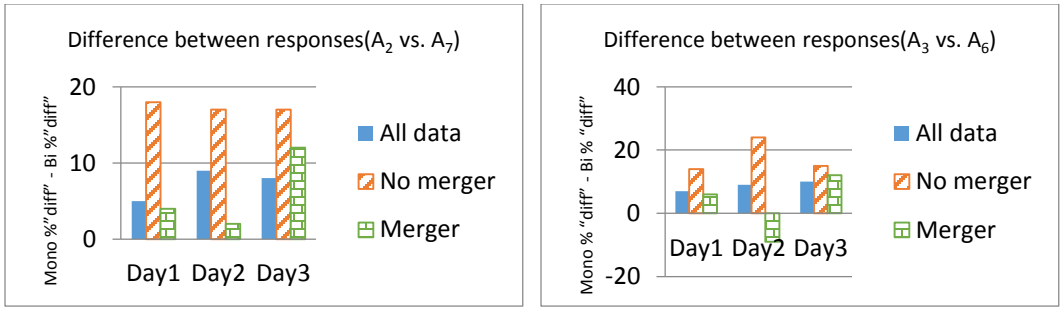

Figure 7. Difference between the percent of "different" responses in the MONOMODAL NOSAMEENVIRONMENTS and BIMODAL SAMEENVIRONMENTS groups, for (a) A_2 vs. A_7 and (b) A_3 vs. A_6.

Of the 31 participants in the BIMODAL SAMEENVIRONMENTS group who answered the questionnaire, 19 claimed to pronounce *cot* and *caught* differently, and 9 claimed to pronounce them the same (3 were unsure). Of the 30 participants in the MONOMODAL NOSAMEENVIRONMENTS group who answered the questionnaire, 14 claimed to pronounce *cot* and *caught* differently, and 13 claimed to pronounce them the same (3 were unsure). Figure 7 shows the results of the difference between the percent of "different" responses between the two groups, for A_2 vs. A_7 and A_3 vs. A_6. Results were not significant, but it can be seen that there are trends for a larger contrast between the two groups in those participants without the merger. This suggests that the difference found between the two groups' vowel results cannot be simply attributed to a greater proportion of people with the *cot-caught* merger in the BIMODAL SAMEENVIRONMENTS group.

6.2 Analysis over Time

No evidence is found for either the Frequency Cue or the Lexical Cue becoming stronger relative to the other over time. If it were the case that the Frequency Cue became stronger over time, 1) the percent of "different" responses in the BIMODAL SAMEENVIRONMENTS group should increase over time, and 2) the percent of "different" responses in the MONOMODAL NOSAMEENVIRONMENTS group should decrease over time. If it were the case that the Lexical Cue became stronger over time, 1) the percent of "different" responses in the BIMODAL SAMEENVIRONMENTS group should decrease over time, and 2) the percent of "different" responses in the MONOMODAL NOSAMEENVIRONMENTS group should increase over time. As can be seen in Figure 8, no evidence is found for either of these cases. For the case of vowels, participants answered "different" more over time, regardless of which group they were in. For the case of consonants, no clear trend is observed.

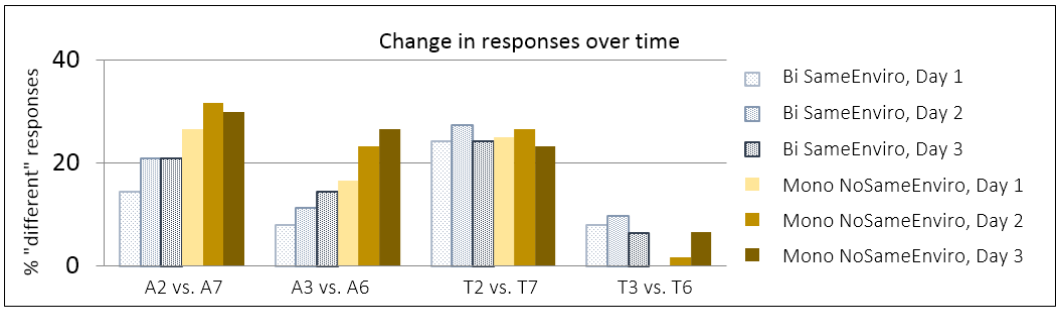

Figure 8. Change in responses over time. For vowel categories, participants answer "different" more often over time; for consonant categories, participant responses show no clear trend.

7 Discussion

To summarize, the vowel stimuli showed clear and consistent, but non-significant trends across all three days and across both the near (A_3 vs. A_6) and far contrast (A_2 vs. A_7). Specifically, there was a consistent trend for the MONOMODAL NOSAMEENVIRONMENTS group to answer "different" more often than the BIMODAL SAMEENVIRONMENTS. On the other hand, the consonant stimuli (T_3 vs.T_6 and T_2 vs. T_7) did not show consistent trends across all three days, and also failed to show consistent tends across the near and far contrasts.

It is important to note though that the Lexical Cue and Frequency Cue are matters of degree. For example, the Frequency Cue could be strengthened with sharper frequency peaks, and the Lexical Cue could be strengthened with more memorable and/or longer words. The difference found between consonants and vowels may be a result of the experiment design (e.g., it could be due to sharper perceived boundaries in the consonant stimuli chosen for this experiment, possibly leading to different "cue strengths"), or to an actual difference in how learners acquire different types of phonemes. Although acquisition reviews tend to treat phoneme acquisition of all phonemes equally, it is suggested here that the acquisition of different *types* of phonemes be more closely examined in future research.

7.1 Change over Time

No evidence was found suggesting that either cue becomes stronger relative to the other over time of exposure. This finding was considered surprising, given that more repetitions of words should lead to a stronger lexical entry, which in turn should result in a stronger Lexical Cue. In addition, it was thought that a period of sleep in between days would lead to what is known as **lexical consolidation** (Leach and Samuel 2007, Gaskell and Dumay 2003). According to a complementary systems model of lexical learning, the phonetic form of a word can be learned quickly ("lexical engagement", but the integration of this form with existing information requires longer exposure and a period of sleep ("lexical consolidation"). For example, Gaskell and Dumay (2003, 2007) found that newly-learned words did not exhibit lexical competition (that is, an example of integration with existing information) unless participants had slept in between training and testing.

Since the Lexical Cue is a form of integrating newly-learned words with existing information (this is, phonemes), it was thought that the Lexical Cue would become stronger over time. It may be the case that the three-day span for this study was not wide enough to find any effect of lexical consolidation, or this may indicate that the Lexical Cue only requires lexical engagement.

8 Conclusion

This study looked at the early stages of phoneme acquisition. In particular, this experiment examined the interaction of two cues proposed to aid language learners in discovering phonemes: the Lexical Cue and the Frequency Cue (aka. "distributional learning"). Although these two cues are not presented as competing hypotheses, the Frequency Cue is considered by many acquisitionists to be the dominant cue (e.g., see Kuhl, 2004; Gervain and Mehler, 2010; Diehl et al., 2004). This

study suggests that the acquisition of phonemes may depend on the type of phoneme being acquired.

References

Bion, Ricardo, Kouki Miyazawa, Hideaki Kikuchi, and Reiko Mazuka. 2013. Learning phonemic vowel length from naturalistic recordings of Japanese infant-directed speech. *PLoS ONE* 8(2): e51594.

Boersma, Paul, and David Weenink. 2013. Praat: doing phonetics by computer [Computer program]. Vers. 5.3.56. `http://www.praat.org/`.

Crump, Matthew, John McDonnell, and Todd Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE* 8(3): e57410.

Dumay, Nicholas, and Gareth Gaskell. 2007. Sleep-associated changes in the mental representation of spoken words. *Psychological Science* 18(1): 35-39.

Feldman, Naomi, Emily Myers, Katherine White, Thomas Griffiths, and James Morgan. 2011. Learners use word-level statistics in phonetic category acquisition. Paper presented at BUCLD 35, Boston University.

Feldman, Naomi, Emily Myers, Katherine White, Thomas Griffiths, and James Morgan. 2013. Word-level information influences phonetic learning in adults and infants. *Cognition* 127(3): 427-438.

Feldman, Naomi, Thomas Griffiths, and James Morgan. 2009. Learning phonetic categories by learning a lexicon. Paper presented at the 31st Annual Conference of the Cognitive Science Society.

Gaskell, Gareth, and Nicholas Dumay. 2003. Lexical competition and the acquisition of novel words. *Cognition* 89(2): 105-132.

Gervain, Judit, and Jacques Mehler. 2010. Speech perception and language acquisition in the first year of life. *Annual Review of Psychology* 61: 191-218.

Gulian, Margarita, Paola Escudero, and Paul Boersma. 2007. Supervision hampers distributional learning of vowel contrasts. Paper presented at the 15th International Congress of Phonetic Sciences.

Hayes-Harb, Rachel. 2007. Lexical and statistical evidence in the acquisition of second language phonemes. *Second Language Research* 23(1): 65-94.

Klatt, Dennis, and Laura Klatt. 1990. Analysis, synthesis and perception of voice quality variations among male and female talkers. *Journal of the Acoustical Society of America* 87: 820-856.

Kuhl, Patricia. 2004. Early language acquisition: Cracking the speech code. *Nature Reviews* 5: 831-843.

Kuhl, Patricia, Karen Williams, Francisco Lacerda, Kenneth N Stevens, and Bjorn Lindblom. 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* 255: 606-608.

Kuhl, Patricia, and James Miller. 1975. Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science* 190 (4209): 69-75.

Leach, Laura, and Arthur Samuel. 2007. Lexical configuration and lexical engagement: When adults learn new words. *Cognitive Psychology* 55(4): 306-353.

Lisker, Leigh, and Arthur Abramson. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. *Word* 20(3): 384-422.

Maye, Jessica, and LouAnn Gerken. 2000. Learning phonemes without minimal pairs." Paper presented at BUCLD 24, Boston University.

Maye, Jessica, Janet Werker, and LouAnn Gerken. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82: B101-B111.

Nespor, Marina, Marcela Peña, and Jacques Mehler. 2003. On the different roles of vowels and consonants in speech processing and langauge acquisition. *Lingue e Linguaggio* 2(2): 203-229.

Stager, Christine L, and Janet F Werker. 1997. Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature* 388: 381-382.

Swingley, Daniel. 2009. Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society Biological Sciences* 364: 3617-3632.

Thiessen, Erik. 2007. The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language* 56: 16-34.

Vaux, Bert, and Scott Golder. 2003. *The Harvard Dialect Survey*. Cambridge, Mass.

Weenink, David. 2009. The KlattGrid speech synthesizer. Presented at Interspeech, Brighton.2059–2062.

Werker, Janet, and Richard Tees. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development* 7: 49-63.

Department of Linguistics
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3155
moeng@live.unc.edu