Interactions between word and speech sound categorization in language acquisition

by

Naomi Hannah Feldman

B. A., University of Chicago, 2003

Mag. Phil., University of Vienna, 2005

Submitted in partial fulfillment of the requirements

for the Degree of Doctor of Philosophy in the

Department of Cognitive, Linguistic, and Psychological Sciences at Brown University

Providence, Rhode Island

May 2011

This dissertation by Naomi Hannah Feldman is accepted in its present form by the Department of Cognitive, Linguistic, and Psychological Sciences as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date _____ _____
James Morgan, Director

Recommended to the Graduate Council

Date _____ _____
Thomas Griffiths, Reader
(University of California, Berkeley)

Date _____ _____
Sheila Blumstein, Reader

Approved by the Graduate Council

Date _____ _____
Peter Weber
Dean of the Graduate School

Acknowledgments

I am fortunate to have had the guidance of three extremely supportive and dedicated mentors, Jim Morgan, Tom Griffiths, and Sheila Blumstein, who served on my committee throughout graduate school. Although Jim was my official advisor, the distinction between advisor and committee was very much blurred, and all three had important roles in supervising this research. Jim gave me the support and freedom to pursue research projects that matched my interests, while always encouraging me to apply computational techniques to real questions in language acquisition. Despite this freedom, I have a sneaking suspicion that he planned for me to write a dissertation on this topic all along. I remember him asking me during my first year, "What part of language do you think infants acquire first?" (My response at the time was, "I think they must learn phonetic categories first." Boy, was I wrong!) I benefitted many times from his willingness to come in over the weekend for practice talks and from his uncanny ability to take a first draft of a paper, abstract, talk, or grant proposal, tear it apart, and rewrite it in just the right way. Tom taught me nearly everything I know about statistical modeling, including theorems like Bayes' rule ($p(h|d) \propto p(d|h)p(h)$) and Griffiths' rule ($p(\text{PhD}|\text{Tom}) > p(\text{PhD})$). (This last rule is written on a whiteboard in a student office in Tom's lab in Berkeley; we're not sure who first wrote it there, but it's been there for over two years.) I am grateful to him for advising my work even after he left Brown, for providing the opportunity to visit Berkeley for a summer while I was developing a dissertation topic, for contributing brilliant new insights and ideas every time I talked to him, and for always remembering to give positive feedback.

Sheila kept me honest about my claims related to linguistic theory, made sure I was on track to meet the department's requirements, and stood up for me every time I needed it. I appreciated that despite being an extremely accomplished researcher and scholar, she would always treat students as intellectual equals, even when discussing research problems that she had thought about for years.

Many other colleagues made large contributions to the success of this work. Katherine White, Sharon Goldwater, and Emily Myers served as terrific role models, friends, and collaborators on various aspects of this project, with Emily in particular giving valuable guidance on the design of Experiment 1. Andy Wallace helped by providing a program for constructing vowel continua, and Joseph Williams shared a derivation he had worked out that was similar to the lexical-distributional model's likelihood function. The experiments would not have been possible without the help of Lori Rolfe, who keeps the lab running almost singlehandedly, and several research assistants (especially Halie Rando and Clara Kliman Silver) who helped recruit participants.

The Department of Cognitive and Linguistic Sciences at Brown was an excellent environment for pursuing graduate study in computational psycholinguistics. Fellow infant lab members Megan Blossom, Erin Conwell, Glenda Molina, Jie Ren, Lori Rolfe, Melanie Soderstrom, Jae Yung Song, Elena Tenenbaum, Jill Thorson, and Katherine White gave feedback on research ideas, sat through countless computational modeling practice talks, and provided a supportive community personally and professionally. I owe special thanks to Elena for being a fantastic academic twin. The the student-run computational modeling reading group also served as a surrogate lab group, and I enjoyed insightful discussions with regular attendees like Dave Buchanan, Liz Chrastil, Adam Darlow, Brad Doll, Micha Elsner, Phil Fernbach, and Dave McClosky. I am especially grateful to Phil for his idea to start the group and to Adam for informal discussions outside of the reading group on all sorts of topics related to computational modeling. Outside of the department, thanks go to members of the computational cognitive science lab, including Josh Abbott, Joe Austerweil, Liz Bonawitz, Daphna Buchsbaum, Kevin Canini, Chris Lucas, Mike Pacer, Anna Rafferty, Florencia Reali, Lei Shi, Joseph Williams, and Jing Xu, for their warm welcome during my visits to Berkeley.

Finally, I am extraordinarily grateful to all the friends and family who kept me going throughout graduate school. Highlights included camping outings, a cappella retreats, maple syrup weekends, and cherry chocolate cake. It would be difficult to come up with an exhaustive list here, but I think you know who you are.

Contents

List of Tables

List of Figures

Chapter 1

The lexical-distributional hypothesis

Which comes first, the speech sound or the word? Infants learning their native language need to extract several levels of structure, including locations of phonetic categories in perceptual space and identities of the words they segment from fluent speech. It is often implicitly assumed that these steps occur sequentially, with infants first learning about the phonetic categories in their language and subsequently using those categories to help them map word tokens onto lexical items. This work examines an alternative hypothesis: infants might simultaneously try to categorize both speech sounds and words, allowing the two learning processes to interact.

Learning speech sound categories from language input is a complex task. Even productions by a single speaker of a specific speech sound in a specific context show substantial acoustic variability, and phonetic categories show even more variability when one considers productions from a range of speakers (Peterson & Barney, 1952; Hillenbrand, Getty, Clark, & Wheeler, 1995). Figure 1.1 shows a map of vowel categories in American English based on data from Hillenbrand et al. (1995). The overlapping distributions suggest that language learners would need to attend carefully to slight differences in pronunciation between sets of vowel tokens in order to distinguish between these categories, while simultaneously ignoring a large degree of within-category variability.

The distributional learning hypothesis (Maye & Gerken, 2000; Maye, Werker, & Gerken, 2002) proposes that learners attend to clusters of exemplars of individual sounds in acoustic space and hypothesize phonetic categories to best match the locations of these clusters. This account is supported by evidence that infants are sensitive to distributional cues at six and eight months, exhibiting better discrimination of stop consonants when the sounds are embedded in a bimodal distribution than when the same sounds are embedded in a unimodal distribution (Maye et al., 2002; Maye, Weiss, & Aslin, 2008). Computational models of phonetic category acquisition have implemented the distributional learning hypothesis, finding the set of Gaussian, or normal, categories that best describes the distribution of sounds in acoustic space (Boer & Kuhl, 2003; Vallabha, McClelland, Pons, Werker, & Amano, 2007; McMurray, Aslin, & Toscano, 2009; Toscano & McMurray, 2010).

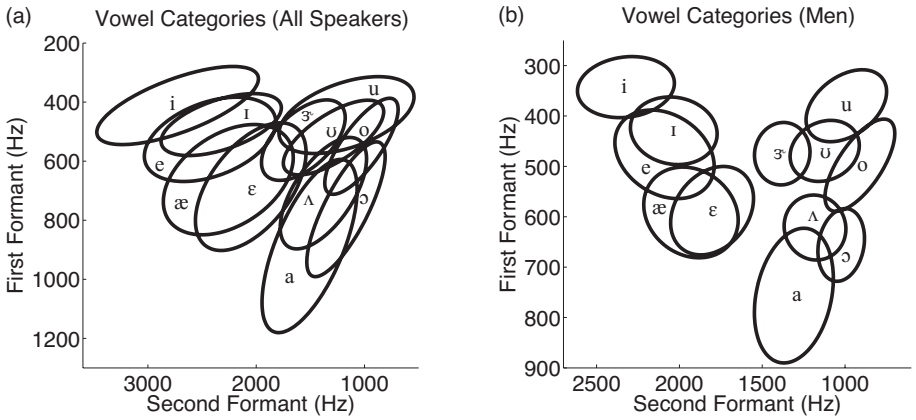

Figure 1.1: Gaussian vowel categories from Hillenbrand et al. (1995) computed based on productions by (a) all speakers and (b) men only. Ellipses delimit the area corresponding to 90% of vowel tokens.

These models have shown promising results in learning well-separated categories. However, simulations in Chapter 4 demonstrate that distributional learning breaks down quickly when categories have a higher degree of overlap because the overall distribution of sounds can appear unimodal, leading learners to underestimate the number of categories. Phonetic category acquisition therefore remains a difficult problem even for a distributional learner.

This dissertation proposes that learners can overcome the problem of overlapping categories by using feedback from their developing lexicon to constrain phonetic category acquisition. Empirical data on infant language acquisition reveal considerable temporal overlap between sound- and word-level learning processes, particularly with respect to word segmentation tasks. Six-month-old infants can use familiar words such as *Mommy* to segment neighboring monosyllablic words from fluent sentences (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005), and a more general ability to segment monosyllabic and bisyllabic words develops over the next several months (Jusczyk & Aslin, 1995; Jusczyk, Houston, & Newsome, 1999), during the same period in which discrimination of non-native contrasts declines. These word segmentation tasks require infants to map word tokens heard in isolation onto word tokens heard in fluent sentences, indicating that infants at this age are performing some sort of rudimentary categorization of the word tokens they segment from fluent speech. This

suggests a learning trajectory in which infants simultaneously learn to categorize sounds and words, with both learning processes occurring between six and twelve months.

The idea of interactive learning is formalized here by constructing a Bayesian "lexical-distributional" model that recovers phonetic categories and lexical items from a pre-segmented corpus. The model's predictions are then tested empirically through experiments with adult participants. Computational and behavioral methodologies provide complementary demonstrations of how information from words and speech sounds might interact during acquisition. Simulations test the extent to which the interaction between words and speech sounds can ultimately help infants acquire their native language, and experiments provide empirical evidence that learners can integrate information from both sources. The two methodologies converge to provide a comprehensive test of the hypothesis that the words infants segment from fluent speech can provide useful constraints to guide phonetic category acquisition.

The dissertation is organized as follows. Chapter 2 reviews previous empirical and computational work on phonetic category acquisition. Chapter 3 describes machine learning techniques fundamental to the lexical-distributional model and reviews previous models of language acquisition that have used these techniques. Chapter 4 introduces the lexical-distributional model and presents simulations to show its improved learning performance over distributional models. Chapter 5 describes experiments investigating whether human learners are sensitive to word-level cues and whether this sensitivity extends to a more realistic language learning situation in which learners need to simultaneously segment words and classify sounds. Finally, Chapter 6 discusses the computational and empirical results in a broader language acquisition framework.

Chapter 2

Background: Speech sound and word categorization

2.1 Introduction

This chapter reviews work on phonetic category acquisition and early word segmentation and categorization. Two types of evidence are considered in each domain. The first type of evidence documents behavioral changes that result from learning that occurs outside of the laboratory, such as patterns of discrimination that change over time to reflect phonetic category knowledge. These types of behavioral changes help elucidate the natural time course of language acquisition. The second type of evidence comes from experiments in which learners are presented with a simplified artificial language and subsequently show evidence of having recovered properties of the artificial language. Positive learning results in this type of experiment indicate sensitivity to the cues contained in the artificial language, providing information about potential mechanisms that learners might use in language acquisition. Although phonetic category learning and word segmentation have typically been considered separately, the last part of this chapter emphasizes research that concerns the relationship between the two learning processes.

2.2 Phonetic category learning

2.2.1 Changes in discrimination

One of the first challenges facing infants during language acquisition is learning which sounds are distinctive in their native language. Studies examining perceptual changes relating to sound category learning have demonstrated that learning in this domain occurs quite early, during the second half of the first year of life. Infants initially show the ability to discriminate contrasts that occur in many different languages (e.g. Eimas, Siqueland, Jusczyk, & Vigorito, 1971; Trehub, 1976; Werker & Tees, 1984). Discrimination of consonants that are contrastive in foreign languages, but not in the native language, declines between six and twelve months (Werker & Tees, 1984). The precise timing and extent of this decline for a particular contrast seems to depend on factors specific to the

sounds in question, such as frequency or markedness (Anderson, Morgan, & White, 2003), relation to native language categories (Best, McRoberts, & Sithole, 1988), or gestural similarity (Best & McRoberts, 2003). The decline is accompanied by enhanced discrimination of native language contrasts (Narayan, Werker, & Beddor, 2010; Kuhl et al., 2006; Sato, Sogabe, & Mazuka, 2010). Many of the studies on perceptual changes in infancy have focused on consonant discrimination ability, as vowel perception remains continuous even into adulthood (Fry, Abramson, Eimas, & Liberman, 1962). However, the available evidence suggests that these perceptual changes begin for vowels even earlier than for consonants. Kuhl, Williams, Lacerda, Stevens, and Lindblom (1992) showed evidence of language specificity in vowel perception in English and Swedish learning infants as early as six months, with perceptual patterns reflecting the locations of vowel prototypes in infants' respective native languages. Spanish learning infants' ability to discrimination similar vowel sounds [e] and [ɛ], contrastive in Catalan but not in Spanish, declines between four and eight months (Bosch & Sebastián-Gallés, 2003), and there is evidence that English-learning infants lose sensitivity to a backness contrast in German high rounded vowels by six to eight months as well (Polka & Werker, 1994). These studies provide evidence that both vowel and consonant perception change during the first year of life to better match the phonetic contrasts that are present in the native language.

These results are generally interpreted as evidence that infants are acquiring native language phonetic categories during the period between six and twelve months, meaning that phonetic category learning mechanisms for extracting category information from linguistic input must be available at an early age. Many different sources of information are potentially useful for category learning, such as the presence of contrastive word forms or the distribution of sounds along phonetic dimensions. Experiments have therefore investigated whether learners are sensitive to these types of cues, especially during the first year of life.

2.2.2 Learning from minimal pairs

In linguistics the contrastiveness of two sounds is often established by identifying minimal pairs, words with different meanings whose forms differ only by a single sound. In the language acquisition literature it has been implicitly assumed that learners use this type of approach as well, using words that differ by a single phoneme, such as *bad* and *bed*, to determine that the minimally different phonemes /æ/ and /ɛ/ belong to different categories. The fact that these words have different meanings indicates that the sound contrast between the two words is meaning-bearing, or phonemic, in the language.

Minimal pair analyses are fundamental to linguistic analyses, helping linguists recover the set of phonemes in a language (Trubetzkoy, 1939). Minimal pair based therapy approaches have also been used successfully in clinical settings with children who have phonological impairments (e.g. Ferrier & Davis, 1973; Gierut, 1989), indicating that preschool aged children can use this type of information to constrain phonemic learning. However, it is not clear that infants know enough word meanings to make use of a minimal pair approach at the time when they are learning phonetic categories. Infants between 13 and 15 months can learn associations between words and objects in laboratory settings (Woodward, Markman, & Fitzsimmons, 1994; Schafer & Plunkett, 1998; Werker, Cohen, Lloyd, Casasola, & Stager, 1998), or even at 12 months under some circumstances (Curtin, 2009), but this is after the period when the bulk of phonetic category learning is typically thought to occur. Infants do know some familiar words at earlier ages. They preferentially attend to their own name as early as 4.5 months (Mandel, Jusczyk, & Pisoni, 1995), correctly associate 'Mommy' and 'Daddy' with the correct parent at six months (Tincoff & Jusczyk, 1999), begin to show comprehension of a few familiar words in naturalistic settings around nine months (Benedict, 1979), and listen longer to lists of common words in the laboratory around 11 months (Hallé & Boysson-Bardies, 1994). However, their receptive lexicons are not likely to be dense enough to support minimal pair based learning at this age (Charles-Luce & Luce, 1995).

Sensitivity to the information contained in minimal pairs is a prerequisite for a minimal pair based learning strategy. Yeung and Werker (2009) explored whether infants are sensitive to minimal pairs at nine months. They tested English-learning infants on a Hindi dental-retroflex contrast for which a decline in discrimination had been shown between six and twelve months. Infants' ability to discriminate these sounds was examined after the infants heard the two sounds presented together with either two consistently different objects, or interchangeably with the same two objects. The authors found evidence of better discrimination in the group that saw the sounds paired with consistently different objects, suggesting that the correlation with visual information helped infants determine that these were different sounds. This provides support for the minimal pair hypothesis by demonstrating early sensitivity to the relevant cues.

However, other studies have yielded results that seem to go against the minimal pair hypothesis. When infants are taught minimal pairs during early word learning, they do not use this information in the way a minimal pair based learning theory would predict. This is especially evident from a series of experiments that tested infants in a word learning task known as the switch task (Stager & Werker, 1997), which tests infants' ability to form associations between words and objects. In the two-word version of the switch task, infants are habituated to two word-object pairings; during test the pairings are switched so that labels are paired with the other familiarized object. In the one-word version of the task, infants are habituated to a single word-object pairing, and during test they see the same object paired with a novel acoustic label. Success on the task is indicated by dishabituation to novel pairings, as indicated by longer looking times.

Stager and Werker (1997) found that 14-month-old infants tested in the switch task failed to dishabituate to novel minimally different labels *bih* and *dih* when those labels were presented together with objects, although they did show the ability to distinguish the sounds when they were not paired with objects. Pater, Stager, and Werker (2004) replicated these negative results with a voicing contrast, place contrast, and two-feature voicing and place contrast. Infants at 14 months succeed in the task when given familiar objects and referents (Fennell & Werker, 2003), and they succeed in

discriminating the same labels when no potential referents are given (Stager & Werker, 1997); the failure seems limited novel labels that are presented together with potential referents. This pattern of results goes against the minimal pair hypothesis, as exposure to a minimal pair with distinct referents fails to help infants distinguish novel words that contain minimally different sounds. In these experiments the presence of distinct referents for similar labels actually decreases infants' tendency to treat sounds from those labels as different, rather than increasing this tendency as the minimal pair account would predict.

This reversal of the minimal pair hypothesis is seen especially clearly in Thiessen (2007). Thiessen replicated the infants' failure in the switch task with labels *taw* and *daw*, then familiarized additional groups of infants to two additional object-label pairs: either *tawgoo* and *dawbow*, or *tawgoo* and *dawgoo*. Infants who heard *tawgoo* and *dawbow* as additional object labels during the habituation period of the Switch task discriminated between daw and taw during test, whereas this facilitation did not occur when the additional object labels were *tawgoo* and *dawgoo*. It was therefore non-minimal pairs, rather than minimal pairs, that allowed infants in this experiment to treat the minimally different syllables as distinct labels. This provides direct evidence against the idea that infants use minimal pairs to separate overlapping phonetic categories in the early stages of word learning.

Older infants succeed at the switch task with minimally different labels (Werker, Fennell, Corcoran, & Stager, 2002), suggesting that the difficulty with minimal pairs is specific to the early stages of word learning. Subsequent research has also shown that 14-month-old infants notice a change in labels when given the same type of familiarization as in the switch task but tested using a different test paradigm, preferential looking, in which two potential referents are provided during test trials (Yoshida, Fennell, Swingley, & Werker, 2009). Thus, minimal pair information does not completely obscure phonemic differences, even in early word learners. However, it is difficult to argue on the basis of these types of results that minimal pairs play an active role in *helping* very young word learners acquire phonemic contrasts.

2.2.3 Distributional learning

Maye and colleagues proposed an alternative to the minimal pair hypothesis, suggesting that learners acquire phonetic categories by attending to specific distributions of speech sounds in the input (Maye & Gerken, 2000; Maye et al., 2002). Under this distributional learning account, learners are hypothesized to obtain information about which sounds are contrastive in their native language based on the distributions of speech sounds they hear. If learners hear a bimodal distribution of sounds along a particular acoustic dimension, they can infer that the language contains two categories along that dimension; conversely, a unimodal distribution provides evidence for a single phonetic category.

Maye and Gerken (2000) tested the effect of distributional information on adults' categorization of prevoiced and short lag stop consonants. Both of these sounds fall into the category of voiced stops in English, but they are contrastive categories in other languages. To test whether these category structures can be learned from distributional information, the authors constructed 8-point voicing continua in which syllable onsets ranged from prevoiced [d] to short lag [t]. Each continuum contained four prevoiced stimuli and four short lag stimuli. Half the participants were familiarized with a unimodal distribution of voice onset times (VOT) and heard tokens 4 and 5 most frequently, mimicking the distribution that would arise from a single phonetic category. The other half were familiarized with a bimodal distribution and heard tokens 2 and 7 most frequently, mimicking the distribution that would arise from two phonetic categories. At test, participants were asked to decide whether the endpoint stimuli belonged to the same category in the language they had just heard. Participants familiarized with the bimodal distribution responded *different* significantly more often than participants familiarized with the unimodal distribution, suggesting that they used the specific distribution of sounds during familiarization to constrain their phonetic categorization.

This result has been extended to vowels by Gulian, Escudero, and Boersma (2007), who familiarized Bulgarian speaking adults with unimodal or bimodal distributions along an /i/-/ɪ/ continuum and an /a/-/ɑ/ continuum. These contrasts are not found in Bulgarian but do appear in Dutch.

Results showed that familiarization with a bimodal distribution facilitated discrimination relative to familiarization with a unimodal distribution. However, this benefit disappeared when participants were given explicit information on category membership during the bimodal familiarization period, suggesting that the statistical learning mechanism may not be compatible with explicit learning processes. This resembles the Stager and Werker (1997) results in that discrimination is worse in the presence of explicit information differentiating the sounds, but differs in that the categorical information from Gulian et al. (2007) did not involve referents.

Adapting the distributional learning experimental paradigm for infants, Maye et al. (2002) familiarized 6- and 8-month-olds with unimodal or bimodal distributions drawn from a [da]-[ta] continuum. During test, infants heard two types of trials: alternating trials, in which the endpoint stimuli 1 and 8 were presented in alternation, or non-alternating trials, in which either stimulus 3 or stimulus 6 was repeated throughout the trial. Infants at both age ranges in the bimodal condition, but not in the unimodal condition, showed a difference in looking times between alternating and non-alternating trials, looking longer at non-alternating stimuli. The authors argued based on these results that the specific distribution of speech sounds that infants heard during familiarization affected their ability to discriminate the endpoint stimuli. Infants in the unimodal condition showed poor discrimination, as would be expected for a within-category contrast, whereas infants in the bimodal condition showed discrimination at the level that might be expected for a between-category contrast. The same type of bimodal familiarization has been shown to facilitate discrimination of a difficult voicing continuum (Maye et al., 2008) and of a place of articulation continuum (Yoshida, Pons, Maye, & Werker, 2010). Although sensitivity to distributional cues remains even after phonetic categories have been acquired (Maye & Gerken, 2000), sensitivity to these cues appears to decrease as acquisition progresses. Ten-month-old infants are able to show enhanced discrimination in response to a bimodal distribution, but only with doubled familiarization time (Yoshida et al., 2010).

These experiments provide evidence that learners are sensitive to distributional cues. To further

test the distributional learning hypothesis, recent computational models of phonetic category learning have investigated whether distributional cues are sufficient to allow learners to recover phonetic categories. These models have assumed that phonetic categories are represented as Gaussian, or normal, distributions of speech sounds and that learners find the set of Gaussian categories that best represents the distribution of speech sounds they hear. Although phonetic categories do not necessarily conform to Gaussian distributions, these types of models are often robust enough to recover clusters of sounds generated from other types of unimodal distributions. If distributional models can recover phonetic categories from realistic input data, this would support the distributional learning hypothesis. However, if these cues are insufficient, this would suggest that supplementary cues are required for successful phonetic category learning.

Boer and Kuhl (2003) used the Expectation Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) to fit a Gaussian mixture model to formant values in mothers' spontaneous productions of *sock*, *sheep*, and *shoe* (containing the /a/, /i/, and /u/ categories). These three vowel categories have relatively large separation in acoustic space, yet simulations showed this to be a difficult clustering problem. When given data from adult-directed speech, the algorithm often produced clusters that were outliers or clusters that overlapped almost entirely. With infant-directed speech the algorithm yielded learning results that were more consistent with category labels, though no quantitative performance measures were given. Even if the algorithm had a reasonable degree of success at learning the three point vowel categories in the latter case, it is not clear whether this would generalize to full natural language vowel inventories.

The EM algorithm is a batch algorithm, finding a locally optimal set of categories based on a set of exemplars given at the start of learning. Two groups of researchers have used an online version of this algorithm whose sequential nature gives more psychological plausibility. The algorithm uses gradient descent to find the means, covariance matrices, and mixing probabilities for the Gaussian categories that give the data maximum likelihood.

McMurray et al. (2009) focused on a voicing contrast, using data that were generated from

Gaussians whose parameters were modeled after production data. The model was successful at acquiring the categories that generated the artificial data, but only when it used maximum likelihood assignments of sounds to categories rather than using the full posterior probability that a sound came from a particular category. Vallabha et al. (2007) applied the same type of algorithm to vowel category acquisition using categories based on laboratory recordings of infant-directed speech from a nonce word elicitation task. Using data from the first and second formant values and duration of each vowel token, the model successfully recovered the /i/, /ɪ/, /e/, and /ɛ/ categories in English and the /i/, /iː/, /e/, and /eː/ categories in Japanese. Vallabha et al. (2007) also created a nonparametric version of their model to address the concern that phonetic categories are typically non-Gaussian; this nonparametric version showed slightly lower performance in learning the Gaussian categories contained in the vowel training data, but was more successful in recovering non-Gaussian distributions in a toy one dimensional example. The simulations in Vallabha et al. (2007) involved neighboring vowel categories, in contrast with the point vowel categories examined by Boer and Kuhl (2003). However, the training data were limited to productions by female speakers in a small set of phonological contexts. Furthermore, the input data were computed and simulations run for each speaker separately, suggesting that the within-category variability presented to the model was less than one would expect to find in real language input.

These computational modeling results are promising indicators that attending to distributional information can help infants acquire phonetic categories. However, the learning problems addressed were simplified compared to the phonetic category learning problem. This is especially true of the vowel category learning simulations, in which categories either had wide separation in acoustic space (Boer & Kuhl, 2003) or excluded substantial within-category variability (Vallabha et al., 2007). Because of these simplifications, it is not clear whether distributional learning can scale up to account for the entirety of phonetic category learning in human language acquisition.

Supplementary cues have been demonstrated to contribute to phonetic category learning. For example, Teinonen, Aslin, Alku, and Csibra (2008) familiarized infants with a unimodal continuum

that ranged from /b/ to /d/ but presented visual stimuli of talking faces pronouncing either /b/ or /d/ alongside these auditory stimuli. One group of infants saw identical visual stimuli for all sounds in the continuum, whereas another group of infants saw one visual stimulus with the /b/-like items and a different visual stimulus with the /d/-like items. Infants in the second group, but not infants in the first group, showed subsequent discrimination of speech sounds from different halves of the continuum. Like the results from Yeung and Werker (2009), this supports a role for visual information in phonetic category learning. More generally, it demonstrates that correlations between acoustic cues and contextual cues can help learners acquire phonetic categories. In previous work this contextual information has taken the form of visual information, but Chapters 4 and 5 examine the contextual information provided by the word token itself.

2.3 Word segmentation and categorization

2.3.1 Segmenting words from natural speech

In the distributional learning studies described above, infants hear only isolated syllables during familiarization. However, young infants between six and twelve months also show evidence of attending to larger units of speech, particularly in word segmentation tasks. Studies using naturalistic stimuli have demonstrated that infants at this age have acquired the ability to segment words from fluent speech using a variety of language-specific cues. As early as six months, infants can segment monosyllabic words that occur next to very familiar words such as "Mommy" or their own name from fluent sentences and map these segmented words onto words heard in isolation (Bortfeld et al., 2005). By 7.5 months, English-learning infants can use stress and other cues to segment certain types of monosyllabic and bisyllabic words (Jusczyk & Aslin, 1995; Jusczyk, Houston, & Newsome, 1999). These infants can segment bisyllabic nouns with a strong-weak stress pattern. However, they do not demonstrate an ability to segment bisyllabic nouns with a weak-strong stress pattern until 10.5 months (Jusczyk, Houston, & Newsome, 1999). Instead, when presented with sentences

containing weak-strong words, where weak syllables reliably follow the target words, 7.5-month-old infants segment strong-weak sequences: presented with a sentence containing *guitar is*, they segment the sequence *taris* rather than the true word *guitar*. This indicates a reliance on stress cues at this age, where strong syllables cue the beginnings of words, mirroring the predominant stress pattern in English. A similar delay in segmentation ability occurs with vowel-initial words, for which segmentation rises above chance performance in the laboratory only at 16 months for nouns (Mattys & Jusczyk, 2001a). This suggests sensitivity to the predominant consonant-initial pattern in English words. Verbs also show a delay relative to nouns, with segmentation beginning at 13.5 months for most types of verbs (Nazzi, Dilley, Jusczyk, Shattuck-Hufnagel, & Jusczyk, 2005); this delay may be related to a variety of factors, including differences in sentence position or prosodic differences between nouns and verbs.

In addition to prosody and syllable onset, infants are sensitive to allophonic cues to word segmentation at 10.5 months. They appear to discriminate between allophones that occur at word boundaries and allophones that occur word-internally, correctly failing to recognize *night rates* in fluent speech when familiarized with *nitrates* (Jusczyk, Hohne, & Bauman, 1999). Learners can use phonotactic cues as well, as demonstrated by 9-month-olds' segmentation of words whose boundaries contain consonant sequences that are common between words, but not words whose the boundaries contain sequences that are common within words (Mattys & Jusczyk, 2001b).

These word segmentation tasks require infants not only to attend to segmentation cues, but also to ignore the within-category variability that distinguishes different word tokens. Thus, infants as young as 7.5 months, who presumably have not yet finished acquiring native language phonetic categories, seem to be performing some sort of rudimentary categorization of the words they segment from fluent speech. The ability to map tokens heard in isolation onto tokens heard in fluent sentences as early as six months suggests that infants are beginning to segment and categorize words at the same time that they are learning phonetic categories. Sensitivity to stress patterns, phonotactics, and allophonic variations all contribute to supporting these early learning abilities. However, these

cues are unlikely to be innate, as prosodic and allophonic patterns differ cross-linguistically; instead the sensitivities may development concurrently with early word segmentation. This leaves open the question of how infants might bootstrap into this system. A large body of research has suggested that sensitivity to statistical regularities provides a potential mechanism for infants to begin segmenting words without prior language specific knowledge.

2.3.2 Statistical learning

Tracking the statistics of sequences of sounds can provide information that is useful for word segmentation. For example, in the sequence "pretty baby", the probability of hearing the syllable *ty* following the syllable *pre* is much higher than the probability of hearing the syllable *ba* following *ty*. This conditional probability that a syllable occurs, given that the preceding syllable has occurred, is referred to in the language acquisition community as a transitional probability.

Saffran, Aslin, and Newport (1996) showed that 8-month-olds track transitional probabilities of the speech they hear, discriminating words from non-words and part-words based purely on this statistical information. Infants were familiarized with a stream of synthesized syllables that consisted of four trisyllabic nonce words in random order. There were no boundaries between successive nonce words, and the only cue to word segmentation was the probabilistic structure of the speech stream. Transitional probabilities between successive syllables within a single word were high, whereas transitional probabilities between successive syllables at word boundaries were much lower. Infants showed a looking time difference when they were subsequently exposed to words from the stream versus groups of syllables that had appeared in the stream but had lower transitional probabilities between syllables. This indicated that the infants were sensitive to the statistical information in the familiarization stimuli. Subsequent studies have repeatedly shown this sensitivity in adults and infants (e.g. Saffran, Newport, & Aslin, 1996; Aslin, Saffran, & Newport, 1998), and even newborns show neural evidence of attending to these types of statistics, showing ERP differences related to syllable predictability (Teinonen, Fellman, Näätänen, Alku, & Huotilainen,

2009).

Infants track statistics in domains other than language (Fiser & Aslin, 2002; Saffran, Johnson, Aslin, & Newport, 1999), indicating domain generality, but there is also evidence that the output from statistical segmentation processes is linguistically relevant. Eight-month-old infants, despite showing a novelty preference for part-words over words in isolation or when the words are embedded in nonsense syllables, instead show a familiarity preference for words over part-words when the words are embedded in English carrier phrases (Saffran, 2001). This familiarity preference resembles that found when real words are embedded in English sentences (Jusczyk & Aslin, 1995), leading the author to hypothesize that infants were treating the segmented words similarly to real words. Familiarization with words embedded in a statistical segmentation task can also facilitate mappings between those words and referents in the switch task (Graf Estes, Evans, Alibali, & Saffran, 2007). This suggests that infants use their sensitivity to transitional probabilities to begin learning potential wordforms for their developing lexicon.

Thiessen and Saffran (2003) showed that when presented with conflicting transitional probability and stress cues, infants transition from a strategy based on transitional probabilities to a stress-based strategy between seven and nine months. Their preference seems to stem from linguistic experience, as familiarization with particular stress patterns can allow infants to use those stress patterns in word segmentation (Thiessen & Saffran, 2007). Taken together with the body of research on word segmentation reviewed above, this suggests a developmental trajectory in which infants first rely on domain general statistical sensitivities, then begin using more language specific strategies like stress or phonotactics. Each of these learning mechanisms, in turn, leads to the word segmentation abilities observed throughout development.

2.4 Combining sounds and words

2.4.1 Phonemes vs. phonetic categories

Psycholinguistic research on phonetic category learning contrasts sharply with much work in phonology. In phonetic category learning, categories are assumed to be Gaussian distributions of sounds, such that the modes of the distribution should roughly correspond to different categories. Context is generally not taken into account, under the assumption that coarticulatory influences should be minimal in comparison to the separation between categories. In contrast, phonological categories can consist of multimodal distributions as long as the sounds in the different modes occur consistently in distinct phonological contexts. Whereas distributional learning can potentially help infants acquire phonetic categories, it does not necessarily lead to the identification of phonemes with these more complex distributions. If one considers phonological categories, then higher level information is almost certainly necessary for acquiring sound categories.

The traditional approach to merging contextually conditioned phonological tokens into a single category involves complementary distribution. Such analyses look at the phonological environments in which these sounds appear and analyze whether the contexts can be characterized by disjoint, and complementary, sets. These analyses typically treat phonological rules or constraints as operating over discrete symbols (e.g. Peperkamp, Le Calvez, Nadal, & Dupoux, 2006; Goldwater & Johnson, 2004), such that the phonetic categories obtained through mechanisms like distributional learning are primitives to a higher-level system. This type of view of the phonetics-phonology interface is articulated explicitly in theories such as that of Keating (1984).

To test whether human learners are sensitive to complementary distributions of phonetic categories, White, Peperkamp, Kirk, and Morgan (2008) examined the extent to which infants could learn to interpret predictable acoustic shifts that were associated with preceding phonemic context as within-category variability. They familiarized 8- and 12-month-old infants with a language in which either stops or fricatives showed voicing assimilation: after the nonce word *rot*, these sounds were

always voiceless (e.g. /p/), whereas after the nonce word *na*, these sounds were always voiced (e.g. /b/). When tested on trials in which pairs of novel /b/- and /p/- initial lexical items were played repeatedly, such as *na boli rot poli*, 8- and 12-month-old infants showed differences in looking times based on whether they had previously heard those sounds in complementary distribution. This indicated sensitivity to the patterns of complementary distribution. Furthermore, 12-month-olds learned to treat isolated minimal pairs, such as *poli boli*, differently even when the conditioning contexts were absent. White and colleagues interpreted these changes as demonstrating that the infants had learned to treat these sounds as part of the same category, despite their acoustic differences. In another experiment, adults given a similar set of familiarization stimuli succeeded only when the alternating nouns were paired with pictures of the same objects, suggesting that they needed additional semantic information in order to interpret these sounds as allophones (Peperkamp, Pettinato, & Dupoux, 2003; Peperkamp & Dupoux, 2007). These differences between infant and adult performance may be either experience- or age-related, but overall the results suggest that distributions of sounds across contexts can provide a primary or secondary cue to category membership in both adults and infants.

Although this type of two-stage model is widely accepted, it fails to explain phenomena like incomplete neutralization, in which phonological rules that seemingly shift sounds from one phonetic category to another fail to produce a distribution of sounds that is identical to the target phonetic category. For example, word-final devoicing processes change voiced stops to voiceless stops at the ends of words, and have been described as neutralizing an underlying voicing contrast; however, underlyingly voiced stops have measurably different acoustic characteristics than the underlyingly voiceless stops (Port & O'Dell, 1985).

Another possibility is that phonetic category learning does not precede phonological learning, but rather accompanies it. This hypothesis is put forward most clearly by Dillon, Dunbar, and Idsardi (submitted), who propose a model in which vowel formants are first shifted linearly to compensate for the influence of a neighboring consonant, then clustered into phonetic categories. Because the

correction for the phonological shift precedes the clustering, the categories found by the phonetic category learning algorithm can be viewed as a type of phonological categories. Although the two types of algorithms differ in their characterization of the learning process, this acoustically based account is similar to the complementary distribution account in that it assumes sounds to have been shifted by context, and therefore looks at context in correcting for these shifts.

Because of their focus on context, either type of learning algorithm would require learners to attend to larger units of speech than individual sounds, and therefore cannot be captured under purely distributional learning algorithms. This suggests that contextual information is critical to learning speech sound categories. However, the type of contextual information required for identifying phonological alternations does not necessarily require knowledge of segmented words, but might come from simple sound cooccurrences computed over the entire speech stream. Thus, it remains an empirical question whether phonological learning precedes or accompanies word learning in human learners.

2.4.2 Phonological specificity of early word forms

Studies that have explicitly addressed both sound and wordform knowledge in early language acquisition have yielded mixed results. By the time infants begin to segment words, some evidence suggests that they can already use their preliminary phonetic category representations in word categorization. For example, Jusczyk and Aslin (1995) found that infants could recognize new tokens of monosyllabic words like *cup* that they had heard during familiarization, and that the same infants correctly failed to recognize onset mispronunciations like *tup*. This indicates that the infants were sensitive to category structure. However, it does not necessarily mean that infants came into the experiment with this category knowledge, as this specificity in recognition might also arise from patterns of variability among the familiarization tokens. This type of phonetic specificity has also been tested in the laboratory using a design that does not provide information about category variability before test. Hallé and Boysson-Bardies (1996) tested 11-month-old French-learning infants'

looking times to lists of correctly pronounced and mispronounced words. In this study they failed to find any evidence that infants could discriminate the two. However, Swingley (2005) found that Dutch-learning 11-month-olds preferred to listen to familiar words over onset mispronunciations of those words. This latter finding suggests that infants may have phonetically detailed representations of familiar words at the point where phonetic discrimination of non-native contrasts declines, but the former failure to find such sensitivity suggests that it may not be as robust as the word segmentation studies in Jusczyk and Aslin (1995) would imply. These mixed results come as late as 11 months, when infants have already gathered a good deal of phonetic category knowledge as measured by their discrimination patterns.

These studies found significant differences between the familiar words and nonsense words, indicating that an overall familiarity with the forms of words in the native language accompanies the decline in discrimination for non-native contrasts. Thus, they provide evidence that infants have some knowledge of both phonetic categories and words during this period, but that the learning processes may not be complete.

2.4.3 Consonants and vowels

Evidence suggests that the relationship between sounds and words may be different for consonants and for vowels. Word segmentation results indicate that consonant-initial words are segmented earlier than vowel-initial words (Mattys & Jusczyk, 2001a; Nazzi et al., 2005). This contrasts with the slightly earlier indication of the acquisition of vowel categories as compared to consonant categories (Kuhl et al., 1992; Werker & Tees, 1984) and suggests that although both types of categories are acquired early, consonants may provide more salient cues to word boundaries.

An asymmetry between consonants and vowels is especially evident in a series of statistical learning experiments by Bonatti, Pea, Nespor, and Mehler (2005). In these experiments, adult participants were familiarized with a speech stream that contained four families of words. In the first experiment, members of a family of words had the same sequence of three consonants but had

a variety of intervening vowels. The sequence of words in the speech stream was arranged such that the transitional probabilities between successive vowels were always 0.5, whether within or between words, but the transitional probabilities between consonants were 1.0 within words and 0.5 between words. Only consonants provided a cue to word boundaries. In the second experiment, members of a word family had a constant sequence of three vowels, but had varying consonants, with vowel transitional probabilities determining word boundaries. Participants discriminated between words and part-words in the first but not the second experiment, suggesting that statistical learning operates over consonants but not over vowels.

In a follow-up experiment, Toro, Nespor, Mehler, and Bonatti (2008) tested adult participants' ability to extract rule-like patterns, such as AAB (cf. Marcus, Vijayan, Rao, & Vishton, 1999), by abstracting across words in a segmentation task. Participants were successful when these patterns were instantiated only in the vowels of the CVCVCV words, but were unsuccessful when only the consonants carried this information. The authors argued based on this pair of results that learners are biased to assume that consonants carry lexical information, whereas vowels carry grammatical information. However, it is clear that at some point learners keep track of which vowels are in particular words, as even young word learners can notice vowel mispronunciations of familiar words (Mani & Plunkett, 2007, 2010).

These differences between consonant and vowel learning suggest the possibility that different statistical mechanisms apply in each case. Consonants are used in transitional probability computations, whereas vowels are used in more abstract pattern learning. If the statistics of vowels and consonants are used in different ways, it is also possible that the respective categories are learned in different ways. Vowel categories typically have more overlap than consonants, and thus potentially present a more difficult learning problem. Whereas distributional learning has shown promising results empirically and computationally in consonant category learning, it is possible that vowel learning does not rely as heavily on these distributional statistics given the high degree of overlap between categories. Here experiments and simulations will test reliance on word-level statistics

specifically for vowel category acquisition.

2.5 Summary

Infants seem to learn a good deal about phonetic categories in their language between six and twelve months, as evidenced by a decline in discrimination of non-native consonant contrasts. Two main hypotheses have been proposed to explain how infants learn these categories, the minimal pair hypothesis and the distributional learning hypothesis. There is evidence that infants use both strategies, at least to some extent, during the second half of their first year of life. However, there is also evidence that argues against the minimal pair hypothesis in infants who are in the early stages of word learning.

Research in phonetic category acquisition has typically ignored the contemporaneous learning process of word segmentation, implicitly assuming either that each process proceeds independently in parallel, or that phonetic category learning is nearly complete by the time word segmentation begins. However, research on word segmentation suggests that infants have access to word-sized units during much of the time when they are learning phonetic categories. They demonstrate abilities to use transitional probabilities, stress, and allophonic cues to segment speech into acoustic word tokens at this age. Furthermore, the word segmentation tasks with natural speech require infants to map words heard in isolation onto words heard in fluent sentences, indicating that infants are categorizing different word tokens into some sort of category that corresponds to a word type. The hypothesis tested in the remaining chapters is that the words infants segment from fluent speech can provide a useful cue for phonetic category learning, allowing learners to achieve better learning performance from the same linguistic data.

Chapter 3

Background: Bayesian models of language acquisition

3.1 Bayesian inference

Bayesian models of learning are concerned with finding the optimal statistical solution to a specific inductive problem faced by an ideal observer. The ideal observer is assumed to come to the task with a comprehensive set of hypotheses about the world, prior biases about how likely each hypothesis is to be true, and a model of what sort of data each hypothesis would predict. Bayes' rule (Equation 3.1) governs the way in which the ideal observer uses observed data to update the probability distribution over hypotheses.

$$p(h|d) = \frac{p(d|h)p(h)}{\sum_{h \in H} p(d|h)p(h)} \tag{3.1}$$

The posterior probability $p(h|d)$ represents the probability that an optimal observer would assign to hypothesis h after observing data d. It is proportional to the likelihood $p(d|h)$ and the prior probability $p(h)$ of the hypothesis. The likelihood gives the probability of observing data d assuming that hypothesis h is correct, and provides a measure of how well the hypothesis fits the observed data. The prior probability does not depend on the data d; it represents the probability that the ideal observer assigned to hypothesis h before any data were observed. Although the prior probability distribution is independent of the currently observed data d, it can depend on previously observed data, as discussed in Section 3.3.

Bayesian inference is useful for working with generative models, which specify a joint probability distribution over things in the world by considering each step in the probabilistic process by which those things were generated. It gives a method for recovering information about latent variables which help generate observed data, but which themselves are not observed. Each model specifies a prior probability distribution over hypotheses, corresponding to the way in which the hidden states of the world were generated, and a probability distribution by which data were generated from those hidden states. The learner recovers the statistically optimal posterior probability distribution over hidden states based on prior beliefs and observed data.

Taking medical diagnosis as an example, a generative model would specify how patients' symptoms are generated: they typically arise from an underlying, but unknown, disease. The model would specify a prior probability distribution that governs the frequency with which any particular disease occurs, $p(disease)$. In addition, it would specify a conditional probability of exhibiting a particular set of symptoms, given that a patient has a particular disease, $p(symptoms|disease)$. Faced with a patient exhibiting a set of symptoms, a doctor would need to work backwards to find the most likely disease that generated those symptoms, computing $p(disease|sympotoms)$. Here the hypotheses are the possible diseases that the patient could have, and the observed data are the symptoms. Bayes' rule gives the mathematically optimal way of doing this type of reverse inference.

In contrast to other types of computational models, such as neural networks, Bayesian models give no account of the mechanism by which an ideal observer arrives at a posterior probability distribution over hypotheses. The models are concerned only with identifying the optimal learning outcome and are potentially compatible with a wide range of algorithms that allow the observer to converge on this optimal outcome.

3.2 Language acquisition as inductive inference

3.2.1 Structure of the language acquisition problem

Language acquisition can be thought of as a statistical inference problem if one assumes that learners need to recover the structure of their language based on the linguistic input they receive. The generative model assumed by this type of framework specifies a prior probability distribution over linguistic structures, representing a sort of probabilistic universal grammar, as well as a model of how corpus data are generated based on the structure of the language. Here the hypotheses consist of the possible linguistic structures and the data are corpus data. The learner takes the role of the ideal observer, computing a posterior probability distribution over linguistic structures given the observed data.

This general approach has been used in several domains in language acquisition, modeling how an ideal learner would acquire word forms (Goldwater, Griffiths, & Johnson, 2009), word-object mappings (Frank, Goodman, & Tenenbaum, 2009; Xu & Tenenbaum, 2007), syntax (Foraker, Regier, Khetarpal, Perfors, & Tenenbaum, 2009; Perfors, Tenenbaum, & Wonnacott, 2010; Pearl & Lidz, 2009; Regier & Gahl, 2004), or semantics (Piantadosi, Goodman, Ellis, & Tenenbaum, 2008; Alishahi & Stevenson, 2008, 2010) from corpus data. Questions about universal grammar and language structure can be investigated by manipulating the prior probability distribution over linguistic structures or the nature of the structures being learned. This section reviews examples of such models, showing how acquisition in each domain maps onto this general inductive structure.

3.2.2 Word segmentation

In the word segmentation task, the learner is concerned with identifying the set of word forms in a language, using as data an unsegmented corpus of child-directed speech. This task differs from word learning in that the hypothesized word forms are not mapped to referents or concepts, but are simply assumed to be stored in the lexicon as phonological forms. Goldwater et al. (2009) built a model of word segmentation that was able to recover lexical items from a phonemically transcribed corpus, with a prior bias toward shorter words and a smaller lexicon. The research compared two types of assumptions that learners might have about how the corpus was generated, manipulating the ideal learner's prior probability distribution to examine how these prior biases affected the learning outcome. A unigram learner assumed that each word was generated independently of its neighbors, whereas a bigram learner assumed dependencies between successive words. Results demonstrated that an ideal learner who comes to the word segmentation task expecting bigram dependencies shows better performance than a learner who expects no dependencies, indicating a role for these types of dependencies. Whereas statistical learning is often thought of in terms of simple transitional probabilities, this work suggests that statistical learning should involve more than separating predictive dependencies within words from a lack of predictive dependencies between

words. Instead, it suggests that these two different types of dependencies both exist in linguistic structure.

3.2.3 Word learning

Another popular domain in which Bayesian models have been applied is word learning, where the underlying linguistic structure is assumed to be a mapping between words and objects and the data consist of correspondences between words present in an utterance and objects present in a scene. Although this encompasses a broad question, various models have tackled different components of the overall problem. Xu and Tenenbaum (2007) focused on how learners delimit the set of objects denoted by a label, deciding between subordinate, basic-level, and superordinate categories. Frank et al. (2009) instead focused on lexical learning, examining how words can be mapped to specific objects when multiple such mappings are possible on the basis of simple cooccurrences.

In Xu and Tenenbaum's (2007) model, word-object pairings were presented unambiguously to the model, such that a single label accompanied a single object. This indicated to the learner that the object belonged to the set denoted by the label. The learner was then assumed to entertain several hypotheses, representing different sets that might correspond to the label. For example, when presented with a dalmatian labeled as 'dog', the learner needs to determine whether the word 'dog' refers to all dalmatians, all dogs, or all animals. The basis of this word learning model is the Bayesian generalization model of Tenenbaum and Griffiths (2001), which introduces the idea of a size principle: it is assumed that in generating the observed data, an object from the relevant set is sampled at random, yielding a likelihood term $p(d|h)$ of $\frac{1}{|h|}$ where $|h|$ is the size of the set corresponding to a given hypothesis h. This gives higher probability to concepts corresponding to small sets. In word learning it allows the model to prefer narrower concepts even when the observed data are consistent with broader concepts, simply because of the likelihood term. Thus, the set of dogs should be given higher probability than the set of all animals as a referent for the word 'dog', even though the data are equally consistent with the correct referent being the set of all animals.

This prediction corresponded well with the judgments made by human language learners for novel words.

A second subcomponent of the word learning problem is cross-situational learning, which involves extracting the correct mappings between words and objects from the many possible pairings that occur in the world. This is a non-trivial problem because entire sentences of words accompany entire scenes of objects, and any of these many word-object pairings can be extracted by an associative learner. Frank et al. (2009) examined the difference between inferences that result from two different generative models. In one generative model, words in an utterance were assumed to be selected probabilistically on the basis of the objects present and the word-object mappings in the lexicon. In a second generative model, words were assumed to be selected probabilistically on the basis of a speaker's referential intentions and the word-object mappings in the lexicon, where referential intentions were selected probabilistically from the objects present in a scene. Results showed that the insertion of the extra variable representing a speaker's intentions improved model performance, allowing the model to avoid spurious word-object correspondences in its hypothesized lexicon.

This result emphasizes the importance of selecting the correct generative model when performing inference. If the learner's assumptions about how observed data are generated under different hypotheses does not match the actual generative process, then even an ideal learner can draw incorrect inferences. Bayesian inference only guarantees optimality when the learner's assumptions about the world are accurate. In this case, the assumption that a speaker's utterances are dependent only on the objects present in the world, without the intervening influence of a speaker's mental state, misleads an associative word learner.

3.2.4 Syntax

Syntax is traditionally a field in which there have been strong arguments made for innate linguistic knowledge (e.g. Chomsky, 1957), and the Bayesian framework has been used in this debate because it provides a way of incorporating prior probability distributions that represent different types of

innate linguistic knowledge. Syntax learning has been formalized as learning a set of production rules that make up the grammar; hypotheses correspond to sets of syntactic rules, with prior probability distributions specifying which sets of rules are most likely.

One debate concerns the hierarchical nature of language. Here a Bayesian model has been used to argue that hierarchical structure has a higher posterior probability than linear structure given corpus data and a domain-general prior bias toward simplicity (Perfors, Tenenbaum, & Regier, 2006). Although the likelihood term is highest when the grammar predicts only those sentences that have occurred in the corpus, assigning high likelihood to linear grammars that simply memorize the corpus, these linear grammars grow quickly into complex hypotheses as more corpus data are encountered. The prior bias toward simplicity then disfavors these linear grammars. Modeling results suggested that a phrase structure grammar emerged as the most probable given a sufficiently sized corpus, although the difficulty of searching the space of grammars precluded definitive conclusions on this issue.

Once a learner has determined that linguistic structure is hierarchical, hypotheses about words' specific syntactic properties can be formulated in terms of these hierarchies. This has allowed models to address learnability arguments made about these specific types of syntactic knowledge, where proponents of innate syntactic knowledge have argued that the correct hypothesis cannot be inferred from corpus data. For example, in response to claims that syntactic principles of anaphoric *one* could not be inferred from corpus data due to a lack of unambiguous evidence (Lidz, Waxman, & Freedman, 2003), Regier and Gahl (2004) used the size principle from Tenenbaum and Griffiths (2001) to argue that children were receiving implicit negative evidence. They argued that although children were receiving ambiguous evidence, they should choose the most narrow hypothesis consistent with the data. Foraker et al. (2009) extended this argument, demonstrating that if a learner has already learned certain syntactic generalizations about how verbs and complements combine, they can acquire the syntactic principles governing anaphoric *one*. However, this domain is another in which the optimal learning outcome depends on the specific model used. Pearl and Lidz (2009) argued

on the basis of a similar model that although a learner with perfect knowledge of certain syntactic information could arrive at the correct semantic hypothesis, the same learner would not arrive at the correct hypothesis when jointly inferring the syntactic and semantic properties of anaphoric *one*. Although this debate has not yet been resolved, Bayesian models have made an important contribution, providing a way to quantitatively investigate the learning outcome under various assumptions about the prior information available to the learner.

3.2.5 Semantics

Acquisition of compositional semantics has also been studied in this framework, where the problem has been formalized in terms of learning mappings between words and lambda calculus expressions (Piantadosi et al., 2008). The general framework of lambda calculus is assumed to be known in advance, and the learner's task is only to learn which specific expressions map to which lexical items. The hypothesis space correspondingly consists of mappings between words, their grammatical categories, and their semantic forms. The prior distribution favors hypotheses that were simple, again implementing a type of domain-general bias. Given data consisting of pairings of sentences and corresponding world contexts, the model was able to acquire both simple and complex semantic forms for words in a toy example.

Semantic content is often thought to arise not just from lexical items, but also from the constructions in which those items appear. Alishahi and Stevenson (2008) built a model of how this construction-specific semantic knowledge can be acquired. In the model, hypotheses correspond to sets of constructions, and a learner needs to determine which constructions exist in the language and which construction a sentence belongs to. The prior distribution used in this model favored small sets of verb constructions, and the likelihood function ensured that instances whose semantic and syntactic features are similar are most likely to be assigned to the same construction. This likelihood function depended on semantic features of the verb, rather than on the identity of the verb. Because of this, simulations showed that the semantics associated with verbs that were frequently used in a

construction become characteristic of the construction itself.

3.3 Hierarchical models

One common question in response to Bayesian models concerns the origin of a learner's prior beliefs. Whereas the prior probability distribution over hypotheses is often specified arbitrarily by the modeler in practice, it is more appealing to think of this distribution as having itself been learned from data. One simple way of incorporating prior experience is to assume an ideal observer's prior beliefs to have been shaped by previous observations relating to the hypotheses in question. For example, in a word learning scenario, having seen a dalmatian labeled *dog* shapes a learner's prior beliefs for the next encounter with the same word. These beliefs would no longer include sets corresponding exclusively to houses, fruits, and so on. The ideal observer uses the posterior distribution from one time step as the prior distribution at the next time step.

While this Bayesian updating of beliefs across multiple time steps illustrates how learners can use previously observed data to constrain their beliefs about specific hypotheses, learning is often thought to be more general than this would seem to imply. After learning about the word *dog*, a learner may have different prior expectations when encountering a new word like *cat*. For example, the learner may think that cats are likely to vary much more in color than in shape. Evidence shows that children learn over the course of development to show a shape bias in word learning: older children preferentially extend a label to an object of the same shape, whereas younger children are more willing to extend a label to other objects that have the same color or texture (Landau, Smith, & Jones, 1988; Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). This indicates the acquisition of general knowledge that cuts across different categories.

Kemp, Perfors, and Tenenbaum (2007) characterized this type of abstract learning through a hierarchical Bayesian model. In this type of model there are multiple hidden variables to be inferred jointly by a learner, and the value of one variable determines the prior distribution on another

variable. This leads learners to consider hypotheses at multiple levels of generality, inferring the parameters at the higher level that constrain inference at the lower level. In the case of word learning, the lower level corresponds to learning about the meaning of a particular word, whereas the higher level corresponds to learning about the variability in meaning that words typically display.

The hierarchical model has parameters for each category that characterize the probability of an item in that category having a particular feature: an item from category i has feature j with probability θ_{ij}. Although these parameters are different for each category, they are assumed to have been drawn from the same prior distribution. This prior distribution is unknown, so that learners infer information about the prior distribution at the same time as they infer the parameters for specific categories. In this case, the prior distribution has two parameters, α and β. The parameter α represents the variability in θ parameters across categories, and the vector β represents the expected mean value of θ across categories. Learners have prior beliefs about the values that α and β might take but are uncertain of their actual values. They update their belief in hypotheses about α and β at the same time that they learn θ for specific categories.

There is some evidence of this type of higher-level learning in other domains of language as well. Wonnacott, Newport, and Tanenhaus (2008) showed empirical evidence that adult learners in an artificial language learning experiment attend to higher-level statistics about whether verbs tend to occur in multiple types of verb frames, in addition to learning information about which specific verbs occur in which frames. They familiarized adult learners with a language containing two verb constructions. Participants heard either a lexicalist language, in which seven verbs occurred in the more frequent construction and one verb occurred in the less frequent construction, or a generalist language, in which all eight verbs occurred $\frac{7}{8}$ of the time in the more frequent construction and $\frac{1}{8}$ of the time in the less frequent construction. Participants' generalization patterns when presented with a novel verb in one frame depended on their familiarization condition, with participants who heard the generalist language much more likely to productively use the verb in the unmodeled frame.

Perfors et al. (2010) attributed the difference between conditions to the fact that participants

had acquired different prior probability distributions over the course of the experiment. Whereas participants in the generalist condition had learned that verbs in the language could be used interchangeably in both frames, participants in the lexicalist condition had learned that verbs in the language each occurred in only one frame. This required a model in which learners update hypotheses at multiple levels, where one set of hypotheses concerned individual verbs' behavior, and the second set of hypotheses concerned the way verbs tended to behave overall. The authors applied the hierarchical model for overhypotheses from Kemp et al. (2007) to the problem of acquiring verb argument constructions. Parallel to the above case, each verb had a parameter θ indicating how often it occurred in each of the two constructions. These were assumed to be sampled from a prior distribution over parameters, where α defines the variability between different verbs in the language and β defines the overall frequency with which verbs are found in either construction. In the experiment above, the mean of the beta distribution was $\frac{7}{8}$ in each language, but in the generalist language verbs were close to this mean whereas in the lexicalist language there was more variability in the behavior of different verbs.

A similar approach has been used to describe the acquisition of verb argument constructions in a usage-based framework. Building on the work by Alishahi and Stevenson (2008), Parisien and Stevenson (2010) built a model in which the categories in question were verb argument constructions, with individual instances of verbs used in particular syntactic and semantic frames categorized into constructions that can be generalized across verbs. Parallel to the model by Perfors et al. (2010), learners in this model retained specific information about the constructions that each verb appeared in, as well as general information about which constructions were common across verbs.

The basic principles of this hierarchical approach can be applied to the word and phonetic category learning problems. When hearing a single token of an unknown word, learners have strong prior expectations about the ways in which that word can vary acoustically. Much of this prior knowledge was obtained through experience with other words in the language. In the lexical-distributional model presented in the next chapter, the prior distribution over lexical items is specified in terms of

a phoneme inventory, which is itself learned from language input. The phonemes in the model can be thought of as overhypotheses that define learners' beliefs about the types of acoustic variability that they should expect to find across different tokens of a lexical item.

3.4 Nonparametric models

Recent Bayesian models of language acquisition outside the domain of speech sounds have used a particular class of nonparametric Bayesian models to specify a learners prior beliefs about linguistic analyses (Goldwater et al., 2009; Johnson, Griffiths, & Goldwater, 2007). These models assume the observed data to have been generated by a set of categories, with the probability of each category being generated from a stochastic process called the Dirichlet process (Ferguson, 1973). The Dirichlet process defines a partition of items into categories. It is defined by two parameters, $DP(\alpha, G_0)$, where α is known as the concentration parameter and G_0 is known as the base measure. The concentration parameter governs roughly the number of categories, and the base measure controls the defining characteristics of each category.

The Dirichlet process can be most intuitively described through a construction known as the Chinese restaurant process, which gets its name through a metaphor with Chinese restaurants that have a seemingly infinite number of tables, each of which seems to have infinite seating capacity. The construction works as follows. A first customer enters the restaurant and sits at an empty table. Each subsequent customer who enters the restaurant sits either at an occupied table, or at the next empty table. The probability of selecting a table that already has customers is proportional to the number of customers already at that table, and the probability of selecting a new table is proportional to a concentration parameter α. Thus, if n_i represents the number of customers seated at any given table i, the conditional probability that the new customer sits at table k, conditioned

on all previous customers, is given by

$$p(k) = \begin{cases} \frac{n_k}{\sum_i n_i + \alpha} & \text{for existing tables} \\ \frac{\alpha}{\sum_i n_i + \alpha} & \text{for a new table} \end{cases} \tag{3.2}$$

Note that the probability of a partition of a set of customers does not depend on the specific order in which the customers entered the restaurant, indicating that the process is exchangeable. The joint probability of N customers in a given partition, with K tables occupied, is

$$\frac{\prod_{i=1}^{K} \alpha \Gamma(n_i)}{\prod_{i=0}^{N-1} (i + \alpha)} \tag{3.3}$$

The Chinese restaurant process can be used to specify a probability distribution on partitions of items into categories. Categories correspond to tables and items in those categories correspond to customers. The probability that a new token belongs to a category is proportional to the number of tokens that have already been assigned to that category, creating a "rich get richer" effect in which new tokens tend to get assigned to existing categories. This favors solutions with fewer categories, implementing a simplicity bias similar to the Minimum Description Length principle that has been used previously in the unsupervised learning of morphology (Goldsmith, 2001), word segmentation (Brent, 1999), and syntax (Chater & Vitányi, 2007). However, there is no fixed limit on the number of categories, with each new token having some probability of being assigned to a new category. This prior probability distribution has been particularly useful in representing categorization of word tokens into lexical items (e.g. Goldwater et al., 2009), as it assigns high probability to solutions in which word frequencies follow a power law distribution.

Extending the restaurant metaphor, one can imagine that each table has a single dish on it. All customers seated at the table are associated with that dish. In a categorization problem the dish corresponds to a set of parameters that define a category, such as the phonemic form of a lexical item. Each dish is sampled from the base measure G_0 of the Dirichlet process. The base measure gives a prior probability distribution over category parameters.

Several language acquisition models have made use of the Dirichlet process as a prior probability

distribution over categories being acquired (Goldwater et al., 2009; Johnson et al., 2007; Alishahi & Stevenson, 2008, 2010). In each case, this prior probability distribution encodes a bias toward fewer categories: fewer words in the lexicon, fewer syntactic subtrees, and fewer types of verb constructions. The base distribution G_0 specifies the prior probability distribution over the parameters associated with each category. For example, these sampled parameters can represent the phonemic form of a word, the set of nodes in a syntactic tree, or the semantic and syntactic properties associated with a verb construction. A likelihood term, specific to the domain being studied, specifies how instances in the corpus are generated based on these parameters.

The Dirichlet process has also been used to create hierarchical models, in which learners can update their beliefs at multiple levels. Teh, Jordan, Beal, and Blei (2006) introduced the hierarchical Dirichlet process by drawing an analogy to a Chinese restaurant franchise. Like in the Chinese restaurant process, the probability of a customer sitting at a table is proportional to the number of customers at that table. However, in the Chinese restaurant franchise there are several restaurants, each of which is governed by the same prior distribution $DP(\alpha, G_0)$. Crucially, G_0 is not known and needs to be inferred from the data.

The probability that a customer sits at table k is given by Equation 3.2, where i ranges over the tables in a particular restaurant. When a new table is created in the restaurant, the dish selected for that table is conditioned on the number of times dishes have appeared on tables across the restaurant franchise. That is, the probability of selecting dish j is

$$p(j) = \begin{cases} \frac{n_j}{\sum_i n_i + \beta} & \text{for existing dishes} \\ \frac{\beta}{\sum_i n_i + \beta} & \text{for a new dish} \end{cases} \tag{3.4}$$

where i now ranges over all dishes in the franchise and β is a concentration parameter that governs the probability of inventing a new dish. If a new dish is created, it is drawn from a base distribution over dishes, specified by the modeler. Because the distribution of dishes is the same across restaurants, a learner can pool knowledge from all restaurants to get a better estimate of the probability of each dish. This is similar to the hierarchical models discussed in Section 3.3 in that parameters for

the prior distribution G_0 are drawn from a higher-level probability distribution that needs to be inferred. In this case, the higher level probability distribution is represented by a Dirichlet process over dishes.

The interactive learning model introduced in Chapter 4 implements a hierarchical nonparametric model similar to the hierarchical Dirichlet process. It adopts the structure of the lexicon from Goldwater et al.'s (2009) unigram model of word segmentation, in which the prior distribution over the lexicon is a Dirichlet process. Tables represent lexical items, customers correspond to word tokens, and dishes on each table give the phonemic form of a lexical item. G_0 in this model is a probability distribution over phonemic forms for lexical items: a geometric distribution over the lengths of lexical items, favoring shorter words, with phonetic categories selected independently for each slot in a word. Unlike the model from Goldwater et al. (2009), however, the model introduced in Chapter 4 needs to learn the parameters of the base measure on lexical items. That is, the model needs to recover the parameters of the phonetic category inventory from which the dishes on tables have been drawn. The phonetic category inventory is assumed to correspond to a Dirichlet process, making this case similar to the hierarchical Dirichlet process described above.

Chapter 4

Lexical-distributional model of phonetic category acquisition

4.1 Introduction

The empirical and computational investigations of distributional learning reviewed in Chapter 2 have shown promising results for categories that have substantial separation. However, overlapping categories, such as those shown in Figure 1.1, can mislead a distributional learner. This occurs because the distribution of sounds in two overlapping categories can appear unimodal (Figure 4.1), leading a distributional learner to assign all the sounds to a single category. Although acoustic overlap between categories may be decreased by attending to additional dimensions, such as duration (Vallabha et al., 2007) and formant trajectories (Hillenbrand et al., 1995), this may not be sufficient to solve the learning problem. Estimating category parameters in multiple dimensions is a more difficult problem, and more data are necessary to achieve the same learning outcome. Furthermore, the data in Figure 1.1 were produced in a single phonological context, whereas in actual speech data there is additional variability that arises through contextual variation, such as coarticulation with neighboring sounds. This additional variability is not reflected in Figure 1.1. Phonetic category learning therefore remains a difficult problem, even for a distributional learner.

Two aspects of previous distributional models suggest that more robust learning is possible. First, the algorithms used by these models are guaranteed to find a solution that is locally optimal but are not guaranteed to find a globally optimal solution. Other search algorithms, such as Markov chain Monte Carlo, can overcome the problem of local minima. Second, previous models of phonetic category learning implicitly assume that infants consider only isolated speech sounds. Attending to larger units, such as words, might allow the model to incorporate additional information that enhances phonetic category learning performance.

Simulations in this chapter show that learners can overcome the problem of overlapping categories by using feedback from their developing lexicon to constrain phonetic category acquisition. Interactive learning is beneficial when phonetic categories occur in distinct lexical contexts. The blue and red categories from Figure 4.1 overlap acoustically when considered in isolation, but an

interactive learner can notice that, for example, the blue sounds occur in the word *milk* and the red sounds occur in the word *game*. These lexical contexts are easily distinguishable on the basis of acoustic information and can be used as a disambiguating cue to phonetic category membership. In contrast to minimal pair approaches, this type of interactive learning does not require meanings or referents to be available to the learner. It requires only that learners use acoustic information to categorize word tokens.

A nonparametric Bayesian framework is used to formalize an interactive learning model that learns to categorize sounds and words simultaneously. Simulations compare the performance of an interactive learning model to the performance of purely distributional models in unsupervised learning of English vowel categories. Results demonstrate that incorporating information from words can lead to more robust phonetic category learning than distributional information alone.

4.2 Model formalization

4.2.1 Overview

Nonparametric Bayesian models are optimally suited to investigate interactive learning, as they have shown promising results in previous models of language acquisition (Goldwater et al., 2009; Goldwater, Griffiths, & Johnson, 2006; Johnson et al., 2007) and can be extended to incorporate learning at multiple levels (Teh et al., 2006). One type of nonparametric Bayesian model, the infinite Gaussian mixture model (IMM) (Rasmussen, 2000), has assumptions that correspond to the distributional learning hypothesis. The IMM assumes that a set of unobserved Gaussian categories generates a set of observed sounds and that sounds are generated independently of their neighbors. The learner needs to recover the set of phonetic categories that generated the observed sounds. The model is formalized using a Dirichlet process, and thus allows a potentially infinite number of Gaussian categories but favors solutions with fewer categories, a simplicity bias which is necessary to prevent the degenerate solution in which each acoustic token is assigned to its own phonetic

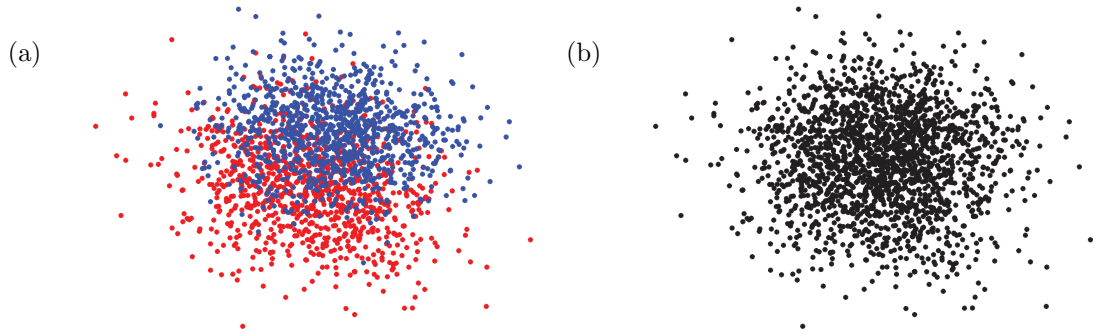

Figure 4.1: (a) Distribution of sounds in two overlapping categories. The points were sampled from the Gaussian distributions representing the /ɪ/ and /e/ categories based on men's productions. (b) These sounds appear as a unimodal distribution when unlabeled, creating a difficult problem for a distributional learner.

category. The strength of the bias toward fewer categories is controlled by a phonetic concentration parameter α_C, with smaller parameter values corresponding to a stronger bias.

The IMM serves as a baseline distributional model and also provides a basis for constructing an interactive lexical-distributional model. The lexical-distributional learning model incorporates the phonetic category structure from the IMM but includes an additional layer of lexical structure. Each lexical item is assumed to consist of a sequence of phonetic categories drawn from the phonetic category inventory. The lexicon is a second Dirichlet process, and the lexical-distributional model is therefore a hierarchical nonparametric model, with the base measure for the lexicon defined in terms of the phonetic category inventory. The model allows a potentially infinite number of lexical items in the lexicon, with a bias toward fewer lexical items whose strength is controlled by a lexical concentration parameter α_L. Again, smaller values of the parameter correspond to a stronger bias. Word tokens in a corpus are assumed to be generated by selecting a lexical item to produce and then drawing an acoustic value from each phonetic category contained in that lexical item. Given a corpus of word tokens, where each token is a sequence of acoustic values, the learner needs to simultaneously recover the set of lexical items and the set of phonetic categories that generated the corpus.

The distributional and lexical-distributional models observe the same corpus data but differ in

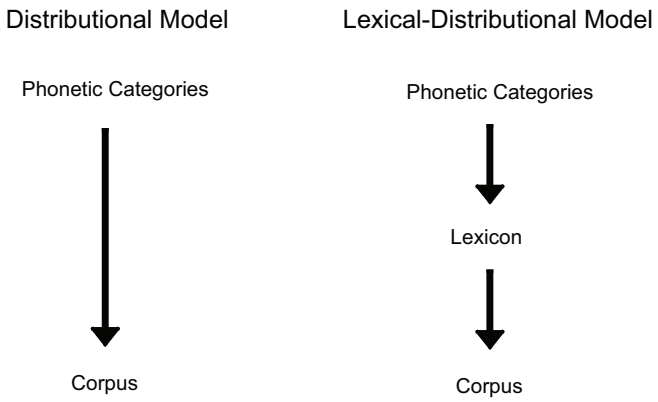

Figure 4.2: Schematic diagram of the distributional and lexical distributional models.

the hypothesis space they assign to the learner (Figure 4.2). The distributional model's hypotheses consist of sets of phonetic categories, and this learner optimizes the phonetic category inventory directly to best explain the sounds that appear in the corpus. In contrast, the lexical-distributional model's hypotheses are conjunctions of sets of phonetic categories and sets of lexical items. This learner optimizes the lexicon to best explain the word tokens in the corpus, and optimizes the phonetic category inventory to best explain the lexical items that are likely to have generated the corpus. This allows the lexical-distributional model to incorporate feedback from the developing lexicon in phonetic category learning.

4.2.2 Lexical-distributional model

The lexical-distributional model is described mathematically in Figure 4.3 and illustrated graphically in Figure 4.4. It assumes a potentially infinite number of categories in a phonetic category inventory, which are combined in sequences to form a potentially infinite number of lexical items. Although the number of phonetic categories and lexical items has no fixed limit, the model assigns higher probability to languages with smaller phonetic category inventories and lexicons, with the strength of this bias governed by the parameters α_C and α_L. Each phonetic category has an associated a Gaussian distribution with a mean and covariance, as well as a frequency of occurrence in the lexicon.

Notation:
μ_c, Σ_c: mean and covariance of phonetic category c
l_k=$(l_{k1}, ..., l_{kn_k})$: lexical item composed of a sequence of phonetic categories
n_k: length of lexical item k
l_{kj}: category for position j in lexical item k; indexes into the phonetic category inventory
w_i=$(w_{i1}, ..., w_{in_{z_i}})$: word composed of a sequence of acoustic values
z_i: category for word i; indexes into the lexicon

Generative model:

$$
\begin{aligned}
& z_i \sim DP(\alpha_L, G_L),\ i = 1..N \\
G_L: \quad & n_k \sim \text{Geom}(g),\ k = 1..\infty \\
& l_{kj} \sim DP(\alpha_C, G_C),\ k = 1..\infty,\ j = 1..n_k \\
G_C: \quad & \Sigma_c \sim IW(\nu_0, \Sigma_0),\ c = 1..\infty \\
& \mu_c \sim N(\mu_0, \tfrac{\Sigma_c}{\nu_0}),\ c = 1..\infty \\
& w_{ij} \sim N(\mu_{l_{z_i j}}, \Sigma_{l_{z_i j}}),\ i = 1..N,\ j = 1..n_{z_i}
\end{aligned}
$$

Figure 4.3: Notation and statistical assumptions for the lexical-distributional model. Word identities in the corpus are drawn from a Dirichlet process whose base measure G_L encodes a geometric prior distribution over word lengths and a second Dirichlet process over phonetic categories that make up each lexical item. This second Dirichlet process, from which phonetic category identities in lexical items are drawn, has a base measure G_C that corresponds to a normal-inverse-Wishart prior over category parameters. Hyperparameters are α_L, α_C, g, μ_0, Σ_0, and ν_0.

Each lexical item has a phonological form, corresponding to a sequence of phonetic categories drawn from the phonetic category inventory, and a frequency of occurrence.

Presented with a corpus consisting of isolated word tokens, each of which consists of a sequence of acoustic values, a learner needs to recover the lexicon and the phonetic category inventory of the language that generated the corpus. To recover samples from the posterior distribution of lexical and phonetic assignments, Gibbs sampling (Geman & Geman, 1984), a form of Markov chain Monte Carlo, is used. Specifically, the simulations use a collapsed Gibbs sampler, integrating out μ_c and Σ_c. The algorithm involves two sweeps, the first to sample category assignments for phonetic category slots in the lexicon, and the second to sample lexical assignments for words in the corpus. In the following algorithm the variables z and l represent the set of word assignments in the corpus and the set of phonetic category assignments in the lexicon, respectively. The variable w represents the set of all acoustic values in the corpus. Minus symbols in subscripts are used to denote the exclusion of particular components; for example, l_{-kj}, is used to denote all phonetic category assignments in l except l_{kj}.

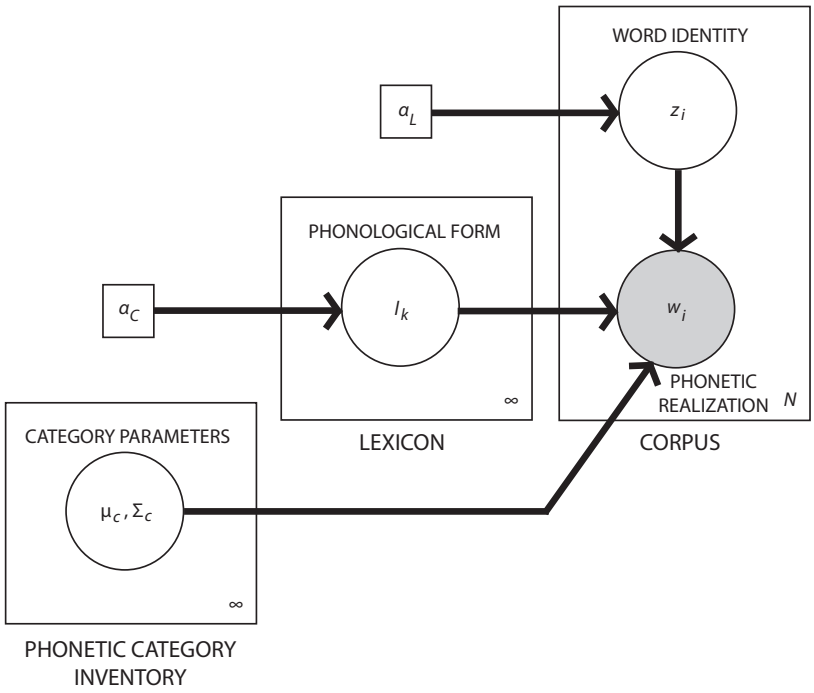

Figure 4.4: Graphical representation of the lexical-distributional model. A phonetic category inventory contains a potentially infinite number of phonetic categories. These categories are organized into a potentially infinite number of lexical items. In generating the corpus, a lexical item is selected for production and acoustic values are drawn from each phonetic category contained in that lexical item. A learner observes the words in the segmented corpus and infers the set of phonetic categories and lexical items that generated the corpus.

In the first sweep, each phonetic category assignment in the lexicon is resampled according to its conditional probability given all other current assignments. If we define w_k as the set of words w_i such that $z_i = k$, this conditional probability distribution can be computed using Bayes' rule as

$$p(l_{kj} = c|w_{kj}, z, w_{-kj}, l_{-kj}) \propto$$
$$p(w_{kj}|l_{kj} = c, z, w_{-kj}, l_{-kj})p(l_{kj} = c|z, w_{-kj}, l_{-kj}) \tag{4.1}$$

The prior distribution $p(l_{kj} = c|z, w_{-kj}, l_{-kj})$ is defined by the Dirichlet process to be

$$\begin{array}{ll} \frac{N_c}{\sum_c N_c + \alpha_C} & \text{for existing categories} \\ \frac{\alpha_C}{\sum_c N_c + \alpha_C} & \text{for a new category} \end{array} \tag{4.2}$$

where N_c is the number of times the phonetic category c has been used previously in the lexicon. The likelihood $p(w_{kj}|l_{kj} = c, z, w_{-kj}, l_{-kj})$ is computed by integrating over all possible means and covariance matrices for the category to obtain the posterior predictive distribution

$$\frac{\Gamma_d(\frac{\nu_c+n}{2})|\Sigma_c|^{\frac{\nu_c}{2}}}{\Gamma_d(\frac{\nu_c}{2})\pi^{\frac{dn}{2}}\frac{n+\nu_c}{\nu_c}^{\frac{d}{2}}}\left|\Sigma_c + \sum_{i=1}^{n}(w_{ij} - \bar{w}_{kj})(w_{ij} - \bar{w}_{kj})^T + \left(\frac{n\nu_c}{n+\nu_c}\right)(\bar{w}_{kj} - \mu_c)(\bar{w}_{kj} - \mu_c)^T\right|^{-\frac{\nu_c+n}{2}} \tag{4.3}$$

where d is the number of dimensions, n is the number of words in the set w_k, and the quantities μ_c, ν_c, and Σ_c encode the current estimates of category parameters based on all sounds assigned to category c. These are defined as

$$\mu_c = \frac{\nu_0}{\nu_0 + n_c}\mu_0 + \frac{n_c}{\nu_0 + n_c}\bar{y} \tag{4.4}$$

$$\nu_c = \nu_0 + n_c \tag{4.5}$$

$$\Sigma_c = \Sigma_0 + \sum(y - \bar{y})(y - \bar{y})^T + \frac{\nu_0 n_c}{\nu_0 + n_c}(\bar{y} - \mu_0)(\bar{y} - \mu_0)^T \tag{4.6}$$

where n_c gives the number of speech sound tokens currently assigned to category c, y are the acoustic values of individual tokens, and $\bar{y}$ represents the mean of those acoustic values. A derivation of this likelihood term is given in Appendix A.

The second sweep reassigns word tokens to lexical items according to Bayes' rule

$$p(z_i = k|w_i, z_{-i}, w_{-i}, l) \propto$$
$$p(w_i|z_i = k, z_{-i}, w_{-i}, l)p(z_i = k|z_{-i}, w_{-i}, l) \tag{4.7}$$

The prior distribution $p(z_i = k|z_{-i}, w_{-i}, l)$ is again given by the Dirichlet process as

$$\begin{array}{ll} \frac{N_k}{\sum_k N_k + \alpha_L} & \text{for existing lexical items} \\ \frac{\alpha_L}{\sum_k N_k + \alpha_L} & \text{for a new lexical item} \end{array} \tag{4.8}$$

where N_k is the number of words in the corpus that have been assigned to lexical item k. The likelihood $p(w_i|z_i = k, z_{-i}, w_{-i}, l)$ for an existing lexical item k is a product of the likelihoods of the speech sounds from each unique category contained in the lexical item, integrating over the parameters of the categories. If we define w_{ic} to be the set of acoustic values in word w_i for which $l_{kj} = c$, this likelihood is

$$p(w_i|z_i = k, z_{-i}, w_{-i}, l) = \prod_c p(w_{ic}|z_i = k, z_{-i}, w_{-i}, l) \tag{4.9}$$

Each term $p(w_{ic}|z_i = k, z_{-i}, w_{-i}, l)$ can be computed using Equation 4.3, replacing the set of acoustic values w_{kj} with the set of acoustic values w_{ic}.

To estimate the likelihood of a new lexical item, a set of 100 samples is drawn from the prior distribution, with the exception that if the word i was previously the only word assigned to a lexical item, that lexical item takes the place of one of the samples from the prior (Neal, 2000). When sampling directly from the prior distribution, each of these 100 samples would receive a pseudo-count of $\frac{\alpha_C}{100}$. In practice, samples are taken only from the portion of the prior distribution for which the likelihood is greater than zero. To correct for this, the pseudo-count of each sample is multiplied by the prior probability of obtaining a word length, syllable template, and set of consonants matching word i.

4.2.3 Distributional models

Two distributional models are used as baseline measures of learning performance. These models illustrate the learning outcomes that are possible in the absence of word-level cues. The first distributional model, the IMM (Rasmussen, 2000), is closely related to the lexical-distributional model. It is defined as a Dirichlet process whose base measure G_C is a normal-inverse-Wishart distribution over Gaussian parameters, identical to that used in the lexical-distributional model (Figure 4.3). Presented with a corpus of sounds, this model finds the set of Gaussian categories that best corresponds to the distribution of sounds in acoustic space.

A collapsed Gibbs sampler is again used for inference, with parameters μ_c and Σ_c integrated out. Sounds are randomly initialized to categories, and each sound in turn is reassigned to a category according to its probability of belonging to that category conditioned on all other assignments,

$$p(z_i = c|w, z_{-i}) \propto p(w_i|z_i = c, w_{-i}, z_{-i})p(z_i = c|z_{-i}) \tag{4.10}$$

where w_i now represents a single sound in the corpus, rather than an entire word, and z_i is the phonetic category assignment of sound w_i. The prior distribution $p(z_i = c|z_{-i})$ is given by Equation 4.2 and the likelihood $p(w_i|z_i = c, w_{-i}, z_{-i})$ is a multivariate t distribution obtained as a special case of Equation 4.3 when $n = 1$,

$$\frac{\Gamma(\frac{\nu_c+1}{2})}{\Gamma(\frac{\nu_c+1-d}{2})}\left|\pi\Sigma_c\left(\frac{\nu_c+1}{\nu_c}\right)\right|^{-\frac{1}{2}}\left(1+(w_{ij}-\mu_c)^T\left[\Sigma_c\left(\frac{\nu_c+1}{\nu_c}\right)\right]^{-1}(w_{ij}-\mu_c)\right)^{-\frac{\nu_c+1}{2}} \tag{4.11}$$

where the parameters μ_c, ν_c, and Σ_c are as defined in Equations 4.4-4.6.

The second baseline distributional learning model, the gradient descent algorithm from Vallabha et al. (2007), enables a direct comparison between results obtained here and results from previous work on phonetic category learning. In contrast to the lexical-distributional model and the IMM, this model does not track category assignments of individual sounds, but rather updates the mean, covariance, and mixing probabilities of each Gaussian in the mixture. A large number of Gaussian distributions are initialized, and upon observing a sound w_i, the posterior probability of category

membership $p(z_i = c|w_i)$ is calculated for each category based on current estimates of category parameters. Each Gaussian category's parameters are then updated as

$$\mu_c \leftarrow \mu_c + \eta(w_i - \mu_c)p(z_i = c|w_i) \tag{4.12}$$

$$\Sigma_c \leftarrow \Sigma_c + \eta(w_i - \mu_c)(w_i - \mu_c)^T p(z_i = c|w_i) \tag{4.13}$$

where η is the learning rate parameter. To update the mixing probabilities, η is added to the mixing probability of the category with highest posterior probability, and the mixing probabilities are renormalized to sum to one.

4.3 Qualitative behavior of an interactive learner

In this section, toy simulations demonstrate how a lexicon can provide disambiguating information about overlapping categories that would be interpreted as a single category by a purely distributional learner. The simulations show that it is not the simple presence of a lexicon, but rather specific disambiguating information within the lexicon, that increases the robustness of category learning in the lexical-distributional learner.

Corpora were constructed for these simulations using four categories labeled A, B, C, and D, whose means are located at -5, -1, 1, and 5 along an arbitrary phonetic dimension (Figure 4.5 (a)). All four categories have a variance of 1. Because the means of categories B and C are so close together, being separated by only two standard deviations, the overall distribution of tokens in these two categories is unimodal. Parameters used for the models were $\alpha_C = \alpha_L = 1$, $\mu_0 = 0$, $\Sigma_0 = 1$, and $\nu_0 = 0.001$; each simulation was run for 500 iterations.

To test the distributional model, 1200 acoustic values were sampled from the four categories, with 400 acoustic values sampled from each of Categories A and D and 200 acoustic values sampled from each of Categories B and C. Results with the IMM indicate that these distributional data are not strong enough to disambiguate categories B and C, leading the model to interpret them as a single category (Figure 4.5 (b)). While this may be due in part to the distributional model's prior

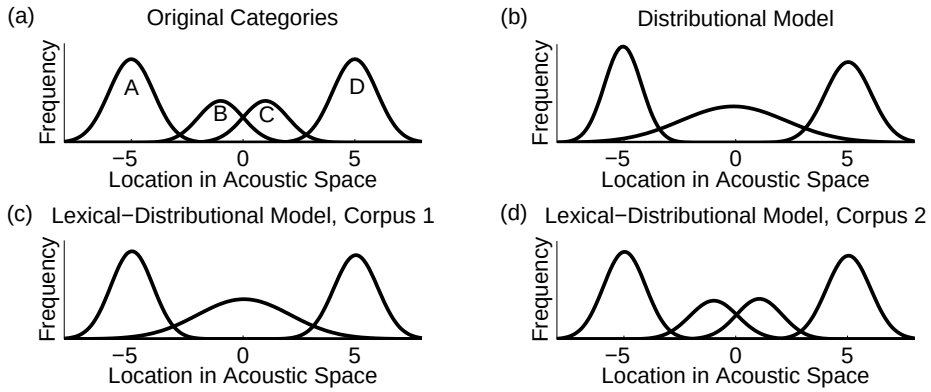

Figure 4.5: Toy data with two overlapping categories as (a) generated, (b) recovered by the distributional model, (c) recovered by the lexical-distributional model from a minimal pair corpus, and (d) recovered by the lexical-distributional model from a corpus without minimal pairs.

bias toward fewer categories, simulations in the next section show that the gradient descent learner from Vallabha et al. (2007), which has no such explicit bias, exhibits similar behavior.

Two toy corpora were constructed for the lexical-distributional model from the same 1200 phonetic values sampled above. The corpora differed from each other only in the distribution of these values across lexical items. The lexicon of the first corpus contained no disambiguating information about speech sounds B and C. It was generated from six lexical items, with identities AB, AC, DB, DC, ADA, and D. Each lexical item was repeated 100 times in the corpus for a total of 600 word tokens. In this corpus, Categories B and C appeared only in minimal pair contexts, since both AB and AC, as well as both DB and DC, were words. As shown in Figure 4.5 (c), the lexical-distributional model merged categories B and C when trained on this corpus. Merging the two categories allowed the model to condense AB and AC into a single lexical item, and the same happened for DB and DC. Because the distribution of these speech sounds in lexical items was identical, lexical information could not help disambiguate the categories.

The second corpus contained disambiguating information about categories B and C. This corpus was identical to the first except that the acoustic values representing the phonemes B and C of words AC and DB were swapped, converting these words into AB and DC, respectively. Thus, the second

corpus contained only four lexical items, AB, DC, ADA, and D, and there were now 200 tokens of words AB and DC. Categories B and C did not appear in minimal pair contexts, as there was a word AB but no word AC, and there was a word DC but no word DB. The lexical-distributional model was able to use the information contained in the lexicon in the second corpus to successfully disambiguate categories B and C (Figure 4.5 (d)). This occurred because the model could categorize words AB and DC as two different lexical items simply by recognizing the difference between categories A and D and could use those lexical classifications to notice small phonetic differences between the second phonemes in these lexical items.

In this model it is non-minimal pairs, rather than minimal pairs, that help the lexical-distributional model disambiguate phonetic categories. Whereas minimal pairs may be useful when a learner knows that two similar sounding tokens have different referents, they pose a problem in this model because the learner hypothesizes that similar sounding tokens represent the same word. The model's behavior resembles that of 15-month-old infants in Thiessen's (2007) experiment, who failed to notice a difference between similar-sounding object labels but were better at discriminating these words when familiarized with non-minimal pairs that contained the same sounds.

4.4 Learning English vowels

4.4.1 Introduction

To conduct a more rigorous test of the model's ability to use word-level information to separate overlapping categories, two simulations were conducted using corpora constructed based on English vowel categories. Vowels show extensive acoustic overlap, and vowel category learning is therefore expected to benefit from lexical information. Simulation 1 illustrates the potential benefit of interactive learning using a lexicon that consists entirely of vowels, whose structure matches the model's assumptions. Simulation 2 tests performance on a lexicon of English words from child-directed speech.

In phonetic categorization the lexical-distributional model is compared with two distributional models, the IMM (Rasmussen, 2000) and the gradient descent algorithm from Vallabha et al. (2007). If lexical information can help separate the distributions associated with English vowel categories, then the lexical-distributional model is expected to outperform these distributional models. In lexical categorization the lexical-distributional model is compared to a baseline model that uses no distributional information from vowels. This baseline model classifies word tokens together if they have the same number of phonemes and, in Simulation 2, if they have the same consonant frame. This comparison quantifies the contribution of distributional information to word categorization.

4.4.2 Methods

Corpus preparation

In each simulation two corpora were used: one with phonetic parameters based on all speakers' productions and the other with parameters based on productions by men only. Each corpus consisted of a sequence of 5,000 word tokens, with word boundaries marked, in which vowel tokens were replaced by acoustic values sampled from the appropriate categories in Figure 1.1. To obtain category parameters, production data from Hillenbrand et al. (1995) were used to compute empirical estimates of category means and covariances in the two-dimensional space given by the first two formant values. The same Gaussian parameters were used to sample each token of a phonetic category that appears in the corpus. The acoustic values in the corpus thus did not reflect any contextual effects, and conformed to the Gaussian assumptions of all the models tested. Consonant tokens in Simulation 2 were represented categorically.

To create each corpus for Simulation 1, a set of lexical items consisting only of vowels, together with a set of lexical and phoneme frequencies, was drawn from the model's prior distribution. Specifically, the set of lexical items was drawn from G_L using a geometric parameter of $g = \frac{1}{3}$. Phonetic category frequencies and lexical frequencies for each corpus were drawn from a Dirichlet

process with $\alpha_L = \alpha_C = 10$. These parameters were chosen to help generate a corpus that contained all twelve vowel categories and were used to sample 5,000 word tokens. These 5,000 word tokens consisted of 22,397 vowel tokens in the corpus based on all speakers' productions and 8,992[1] vowel tokens in the corpus based on men's productions.

Corpora for Simulation 2 were constructed using the CHILDES database parental frequency count (MacWhinney, 2000; Li & Shirai, 2000). Words in the frequency data were converted to their phonemic representations using the CMU pronouncing dictionary. If the dictionary contained multiple phonemic forms for a word, the first was used. Stress markings were removed, and diphthongs were converted to sequences of two phonemes to facilitate their representation in terms of steady state formant values. Any words whose orthographic representation in CHILDES contained symbols other than letters of the alphabet, hyphen, and apostrophe were excluded. In addition, words not found in the CMU pronuncing dictionary were excluded. This resulted in the exclusion of 7,911 types, representing 28,447 tokens (approximately 1% of tokens). This left 15,825 orthographic word types, representing 2,548,494 tokens, whose phonological forms were used to construct the corpus. For each corpus, 5,000 word tokens were sampled randomly with replacement based on their token frequencies. The corpora based on all speakers' productions and men's productions contained 6409 and 6408 vowel tokens, respectively.

Simulation parameters

In the simulations, parameters for the base measure G_C over phonetic categories were $\mu_0 = \begin{bmatrix} 500 \\ 1500 \end{bmatrix}$, $\nu_0 = 1.001$, and $\nu_0\Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. This encoded a weak bias toward the center of vowel space, equivalent to the presence of two points of pseudodata in each category, and was made as weak as

[1] In this corpus, the most frequent lexical item was a single phoneme long, leading to a lower overall number of vowel tokens.

possible while still representing a normalizable prior distribution. The geometric parameter controlling lengths of lexical items was $g = 0.5$. Each simulation used 10,000 iterations of Gibbs sampling with simulated annealing to converge on a sample from the posterior distribution. For Simulation 1 the concentration parameters were set to $\alpha_L = \alpha_C = 1$; using the generating parameters produced qualitatively similar results. For Simulation 2 a range of concentration parameters was tested. In Simulation 2, each phoneme slot in a lexical item was assumed to be designated as a consonant with probability 0.62, the approximate proportion of consonants in the corpus, otherwise it was a vowel. For the purposes of likelihood computation, consonants were assumed to be generated from a Dirichlet process with concentration parameter α_C. This distribution over consonants was assumed to be independent of the distribution over vowels. No attempt was made to optimize these parameters, and they remained constant across simulations.

Parameters used for the gradient descent algorithm were identical to those from Vallabha et al. (2007). A reasonable effort was made to optimize these parameters, and while there was some quantitative improvement with lower standard deviation (cf. McMurray et al., 2009), there was little qualitative change. Results using parameters from Vallabha et al. (2007) are therefore shown to facilitate comparison with previous work.

Quantitative measures

Quantitative measures of model performance were computed over clusterings of vowel tokens (for phonetic categorization) and word tokens (for lexical categorization). The pairwise F-score, defined as the harmonic mean of pairwise accuracy and completeness, measures the extent to which pairs of tokens are correctly assigned to the same category. It ranges between zero and one, with higher scores indicating better performance.[2] To compute the pairwise F-score, pairs of tokens that were correctly placed into the same category were counted as a *hit*; pairs of tokens that were incorrectly assigned to different categories when they should have been in the same category were counted as a *miss*;

[2] A third measure, the adjusted Rand index, gave results similar to the F-score.

and pairs of tokens that were incorrectly assigned to the same category when they should have been in different categories were counted as a *false alarm*. Accuracy (a) was defined as $\frac{\text{hits}}{\text{hits+false alarms}}$ and completeness (c) was defined as $\frac{\text{hits}}{\text{hits+misses}}$. The F-score was computed by taking the harmonic mean of accuracy and completeness, $F = \frac{2*a*c}{a+c}$. The second measure, variation of information (VI) (Meilă, 2007), is an information theoretic measure of the difference between the model's clustering and the true clustering, with lower scores corresponding to better performance. It was computed as $VI(C, C') = 2H(C, C') - H(C) - H(C')$, where H is entropy and C and C' represent the true clustering and the model clustering, respectively.

Lexical categorization performance was analyzed in two different ways for the lexical-distributional model: first by counting each cluster found by the model as a separate lexical item, then by merging all clusters with the same phonemic form into a single lexical item. In both cases, homophones were grouped together in the true lexicon for the purposes of evaluation so that the model would not be penalized for clustering together tokens of words with identical phonological forms such as "they're" and "there".

4.4.3 Simulation 1: Vowel-only lexicon

When tested on the corpora containing vowels sampled from the artificial vowel-only lexicon, the lexical-distributional model recovered the correct set of vowel categories and successfully disambiguated neighboring categories. In contrast, the distributional models mistakenly merged several pairs of neighboring vowel categories (Figure 4.6). Quantitative measures mirrored these qualitative results: F-scores were higher for the lexical-distributional model than for the distributional models, and VI scores were closer to zero for the lexical-distributional model (Table 4.1). The lexical-distributional model also outperformed the baseline word categorization model as measured by number of categories, F-score, and VI (Table 4.2). This indicates that interactive learning improved performance in both the sound and word domains.

These results demonstrate that in a language in which phonetic categories have substantial

		L-D	IMM	GD
All Speakers	Number of categories	**12**	9	4
	Accuracy	**0.985**	0.591	0.399
	Completeness	**0.987**	0.621	0.945
	F-score	**0.986**	0.605	0.561
	Variation of information	**0.163**	2.756	1.956
Men	Number of categories	**12**	9	5
	Accuracy	**0.968**	0.628	0.563
	Completeness	**0.966**	0.594	0.943
	F-score	**0.967**	0.610	0.706
	Variation of information	**0.278**	1.767	1.439

Table 4.1: Phonetic categorization scores for the lexical-distributional model (L-D), the infinite mixture model (IMM), and the gradient descent algorithm (GD) in Simulation 1. The true number of phonetic categories is 12 for each corpus.

		L-D	baseline
All Speakers	Number of categories	**45/40**	11
	Accuracy	**0.988/0.987**	0.652
	Completeness	0.988/0.988	**1.000**
	F-score	**0.988/0.988**	0.789
	Variation of information	**0.315/0.301**	1.145
Men	Number of categories	**50/50**	10
	Accuracy	**0.971/0.971**	0.628
	Completeness	0.969/0.969	**1.000**
	F-score	**0.970/0.970**	0.771
	Variation of information	**0.265/0.265**	1.570

Table 4.2: Lexical categorization scores for the lexical-distributional model (L-D) and the baseline model in Simulation 1. The first number treats each cluster as separate, regardless of phonological form, and the second number treats all clusters with identical phonological forms as belonging to a single lexical item. The true number of lexical items is 42 for the combined corpus and 54 for the men's corpus.

overlap, an interactive system learns more robustly than a purely distributional learner from the same number of datapoints. Positing the presence of a lexicon helps the ideal learner separate overlapping vowel categories, even when phonological forms contained in the lexicon are not given to the learner in advance. However, because Simulation 1 used lexical items drawn from the model's prior distribution, a question remains as to whether the lexicon of a natural language contains enough disambiguating information to separate overlapping vowel categories. Simulation 2 tests this directly using a corpus of lexical items from English child-directed speech.

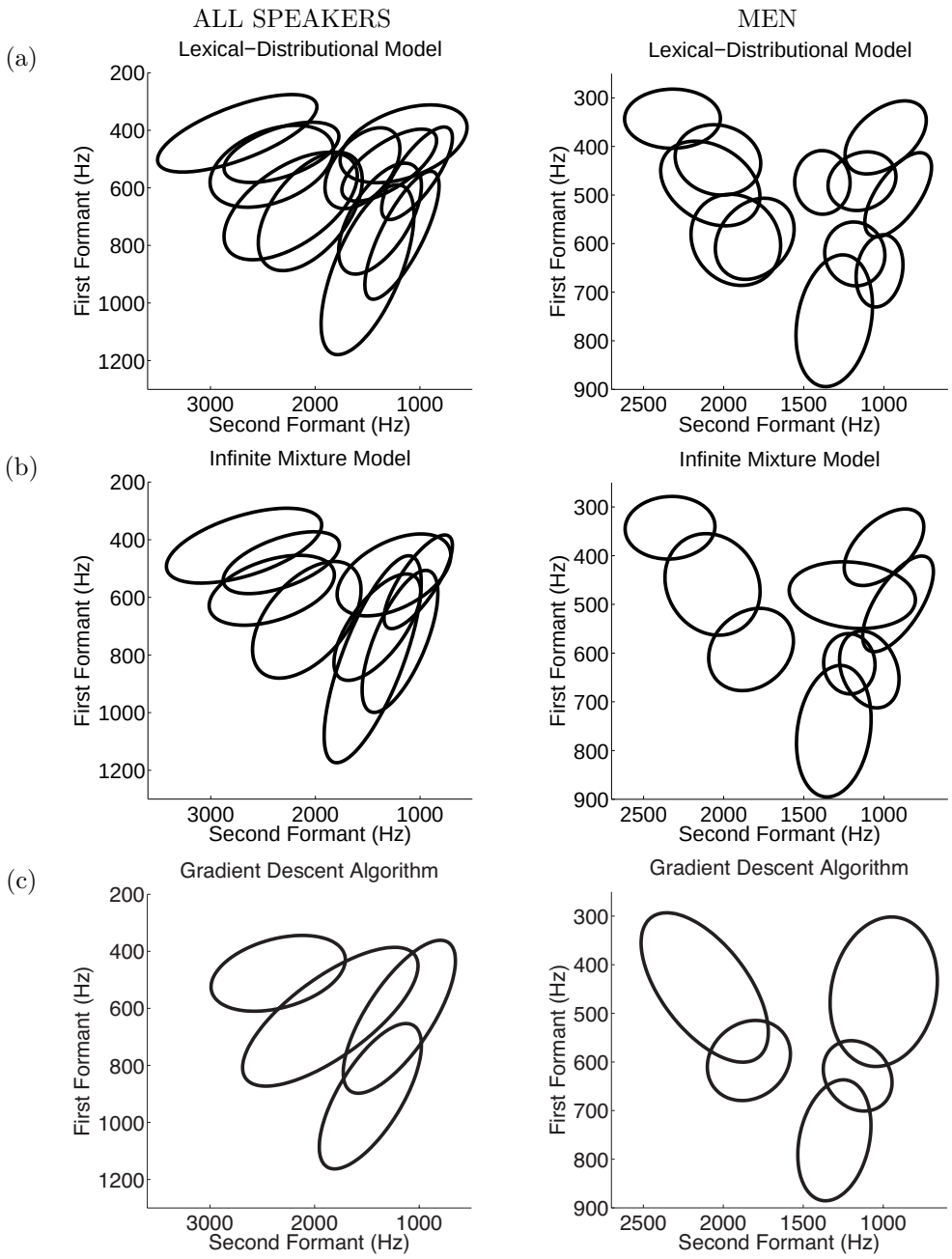

Figure 4.6: Ellipses delimit the area corresponding to 90% of vowel tokens for Gaussian categories recovered in Simulation 1 by (a) the lexical-distributional model, (b) the infinite mixture model, and (c) the gradient descent algorithm.

		L-D					IMM	GD
		$\alpha_L = 1$	$\alpha_L = 10$	$\alpha_L = 100$	$\alpha_L = 1000$	$\alpha_L = 10000$		
All Speakers	$\alpha_C = 0.1$	14	14	13	**12**	**12**	6	4
	$\alpha_C = 1$	14	14	13	**12**	**12**	6	
	$\alpha_C = 10$	14	13	13	**12**	**12**	7	
Men	$\alpha_C = 0.1$	16	14	14	13	**12**	10	5
	$\alpha_C = 1$	17	15	14	13	**12**	10	
	$\alpha_C = 10$	17	16	14	13	**12**	9	

Table 4.3: Number of phonetic categories found by the lexical-distributional model (L-D) and the infinite mixture model (IMM), and the gradient descent algorithm (GD) in Simulation 2. The true number of phonetic categories is 12 for each corpus.

4.4.4 Simulation 2: English lexicon

For simulations with real English lexical items, ranges of parameter values were tested for the phonetic concentration parameter α_C in both the IMM and the lexical-distributional model and the lexical concentration parameter α_L in the lexical-distributional model. This was done to find the parameter values that would perform best on real word and phoneme frequencies and to test the models' robustness to changes in these values. The number of categories recovered by each model is shown in Table 4.3. Performance was quite robust to changes in the phonetic concentration parameter; for this reason, all remaining analyses use a value of $\alpha_C = 10$. Performance of the lexical-distributional model varied depending on the value of the lexical concentration parameter α_L. With a weak bias toward a smaller lexicon, the model recovered the correct set of twelve categories, but with a stronger bias the model hypothesized more than twelve categories. These extra categories had more acoustic variability than the actual categories in the corpus, encompassing more than one vowel category. The distributional models again merged several sets of overlapping categories. Representative examples of each of these types of behavior are illustrated in Figure 4.7. Numerical phonetic categorization performance was consistently higher in the lexical-distributional model than in the distributional models (Figure 4.8 (a)), indicating that even for the models that hypothesized extra categories, information from words improved vowel categorization performance.

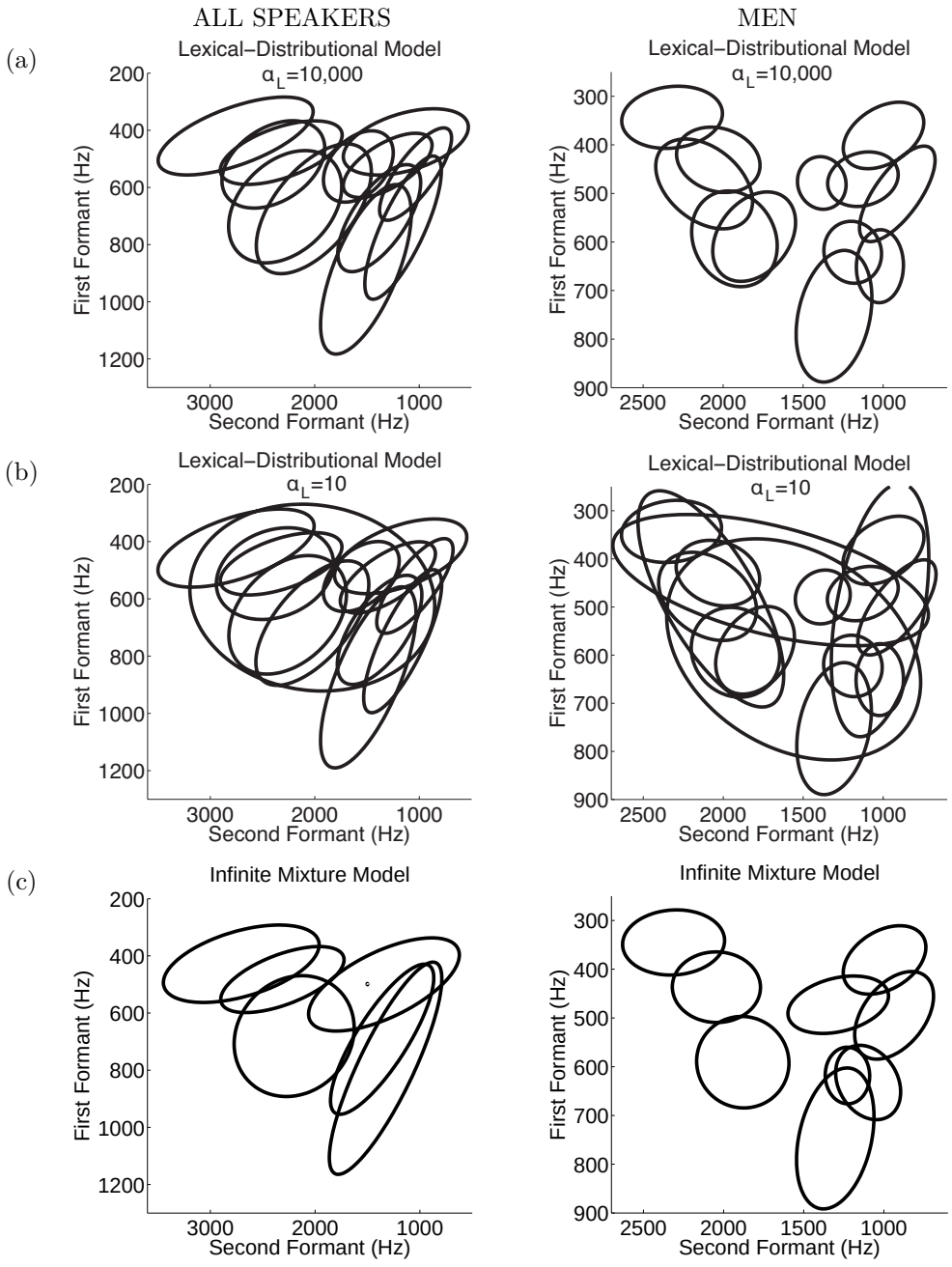

Figure 4.7: Ellipses delimit the area corresponding to 90% of vowel tokens for Gaussian categories recovered in Simulation 2 by (a) the lexical-distributional model with $\alpha_L = 10,000$, (b) the lexical-distributional model with $\alpha_L = 10$, and (c) the infinite mixture model.

		L-D					baseline
		$\alpha_L = 1$	$\alpha_L = 10$	$\alpha_L = 100$	$\alpha_L = 1000$	$\alpha_L = 10000$	
All Speakers	$\alpha_C = 0.1$	900/900	916/915	969/936	1145/1007	1601/1080	852
	$\alpha_C = 1$	899/898	912/910	968/943	1138/995	1605/1078	
	$\alpha_C = 10$	900/899	926/920	958/934	1164/989	1602/1086	
Men	$\alpha_C = 0.1$	912/912	936/934	977/953	1124/1002	1499/1060	840
	$\alpha_C = 1$	905/905	934/928	988/963	1126/1000	1498/1055	
	$\alpha_C = 10$	901/901	933/931	978/957	1117/1002	1502/1057	

Table 4.4: Number of lexical items found by the lexical-distributional model (L-D) and the baseline model in Simulation 2. The first number treats each cluster as separate, regardless of phonological form, and the second number treats all clusters with identical phonological forms as belonging to a single lexical item. The true number of lexical items is 1019 for each corpus.

The number of lexical items recovered by each model is shown in Table 4.4. Each lexical-distributional model recovered more lexical items than the baseline model, indicating that the model used distributional information to separate distinct lexical items that had the same consonant frame. As predicted, stronger biases toward a smaller lexicon resulted in the recovery of a smaller lexicon. With a strong bias, the model recovered fewer lexical items than were used to generate the corpus, merging items that should have been separated. With a weak bias, the model recovered more lexical items than were used to generate the corpus, separating items that should have been assigned to a single category. This separation of items that should be categorized together decreased quantitative lexical categorization performance, but quantitative measures were higher when different clusters with the same phonemic form were treated as a single lexical item for evaluation (Figure 4.8 (b)).

The merger of lexical items in models with a strong lexical bias is related to the extra categories hypothesized by these models. A representative example of the content of an extra category is shown in Figure 4.9. The category consists largely of minimal pairs, words in which all but one phoneme are identical, that are assigned by the model to a single lexical item. The extra category captures the fact that the distribution of acoustic values in these merged lexical items does not fit any of the existing twelve vowel categories, but instead has a broader distribution.

In summary, the lexical-distributional model consistently outperformed the distributional models in phonetic categorization performance, indicating that words in the English lexicon contain useful

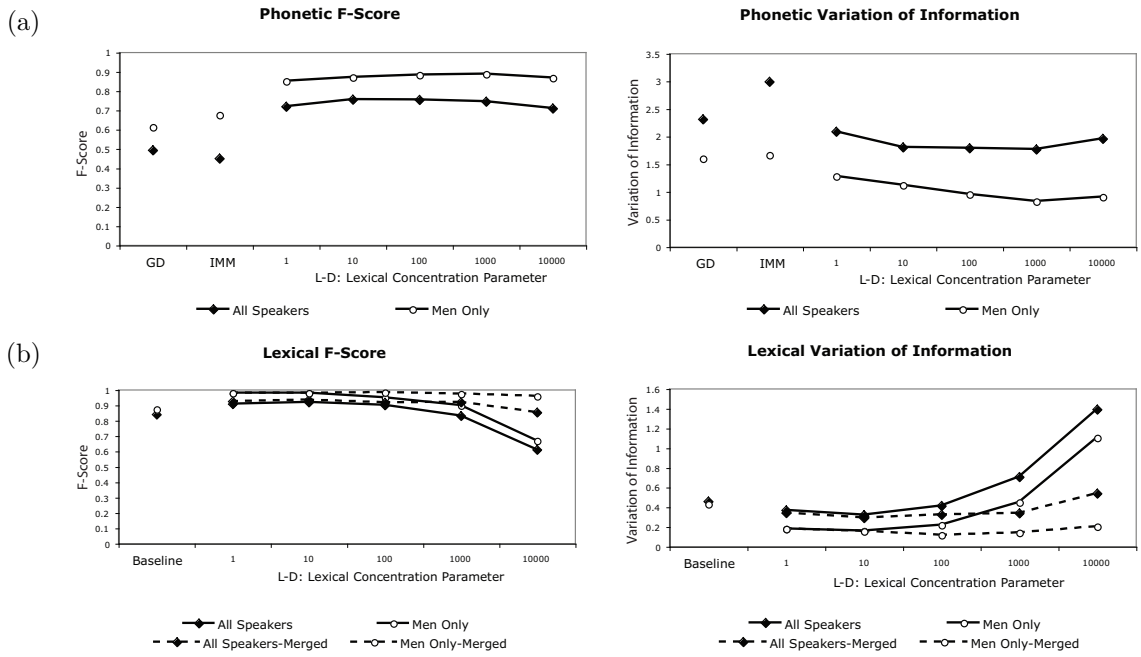

Figure 4.8: (a) F-score and variation of information measuring phonetic categorization performance by the gradient descent algorithm (GD), infinite mixture model (IMM), and lexical-distributional model (L-D). The lexical-distributional models consistently outperform the distributional models, indicating that information from words provides a useful constraint for phonetic category learning. (b) F-score and variation of information measuring lexical categorization performance by the baseline model and lexical-distributional model. Solid lines treat each cluster in the lexicon as its own lexical item, whereas dotted lines treat all clusters with the same phonemic form as a single lexical item. The lexical-distributional models with low concentration parameters outperform the baseline model under both metrics, whereas models with high lexical concentration parameters perform well only when evaluated based on phonemic form.

b**a**t, b**i**t, b**oa**t
b**ea**n, b**ee**n, b**o**ne
bedr**oo**m
bicyc**le**
br**ea**k, br**o**ke
check**i**ng
dann**y**, dinn**er**, donn**a**
d**ir**ty
ev**er**y
f**i**gure
f**i**t, f**oo**t
gr**a**pe, gr**ou**p
happ**y**, hipp**o**
h**a**ts, h**ur**ts
map**le**
pl**ay**ing
p**o**lka
r**ea**l, r**o**ll
s**ee**saw
t**e**nts
t**i**ssues
tomat**oe**s
w**a**ke, w**ee**k, w**or**k
w**a**lks, w**ee**ks
w**ay**, w**ere**, wh**oa**
w**i**nd, w**ou**nd

Figure 4.9: Contents of one of the super-categories shown in Figure 4.7 (b). The sounds identified as belonging to the super-category are highlighted in bold. Multiple orthographic forms are listed next to each other if tokens of that lexical item correspond to more than one word. Many of these lexical items are minimal pairs that the model mistakenly categorizes together.

information for phonetic category learning. With a strong bias toward a smaller lexicon, the model showed high lexical categorization performance but hypothesized extra phonetic categories to account for the high acoustic variability that resulted from erroneously merging minimally different words. With a weaker bias, the model's lexical categorization performance decreased because lexical items were split into multiple categories, but this allowed the model to find the correct set of categories.

4.5 Discussion

This chapter introduced a model of phonetic category acquisition that allows interaction between speech sound and word categorization. The model was not given a lexicon a priori, but was allowed to begin learning a lexicon from the data at the same time that it was learning to categorize individual speech sounds, allowing it to take into account the distribution of speech sounds in words. This lexical-distributional learner outperformed a purely distributional learner on several corpora whose categories were based on English vowel categories, showing better disambiguation of overlapping categories from the same number of data points.

Whereas distributional models assume that learners are focused on recovering phonetic categories from patterns of acoustic variability, the lexical-distributional model shifts the focus of learning to the word level. It treats phonetic categories as a higher-level hypothesis about the type of variability that learners should expect to find in lexical items. The learner selects the phoneme inventory to best explain the variability that occurs in hypothesized lexical items rather than the overall variability in speech sounds. In this framework, knowledge of phonetic categories is a form of Bayesian overhypothesis (Kemp et al., 2007; Perfors et al., 2010), benefitting learners by allowing accurate prediction of what sort of variability a new lexical item is likely to exhibit on the basis of only a few acoustic tokens.

The results from Simulation 2 concerning minimal pairs are striking because linguistic analyses

often cite minimal pairs as evidence that similar sounds belong to different categories. In this model such pairs are treated in the opposite way, being assigned to a single phonetic category and even a single lexical item. Minimal pairs only help distinguish similar sounds if referents are available to the learner; their role in phonetic category acquisition therefore critically depends on the extent to which young infants can make use of associations between form and meaning to separate acoustically overlapping categories. Empirical data suggest that even in the early stages of word learning, learners have trouble using these types of associations in certain laboratory tasks to separate minimal pairs (Stager & Werker, 1997), but performance improves when disambiguating lexical contexts are provided (Thiessen, 2007).

The lexical-distributional model can overcome the problem of minimal pairs when given a weaker bias toward a smaller lexicon. However, real learners may be able to overcome this problem in other ways. For example, infants have knowledge of common speech sound sequences by nine months (Jusczyk, Luce, & Charles-Luce, 1994). Knowledge of phonotactics would likely improve performance by assigning higher probability to phoneme sequences that occur in the lexicon, raising the probability of generating the same consonant frame twice. In addition, human learners have access to semantic information about words that can potentially help them separate minimal pairs in the lexicon. Though most mappings between words and objects are thought to be learned later in development (e.g. Woodward et al., 1994; Werker et al., 1998), 9-month-old infants can use cooccurences between sounds and objects to constrain phonetic category learning (Yeung & Werker, 2009). Although such information is not likely to be robust enough to support a purely minimal pair based learning strategy, it can potentially help learners combine words that have been assigned to different categories or disambiguate similar-sounding words that have erroneously been merged.

In generalizing these results to more realistic learning situations, it is important to take note of a simplifying assumption that was present in the model: Speech sounds in phonetic categories were assumed to follow the same Gaussian distribution regardless of phonetic or lexical context. In actual

speech data, acoustic characteristics of sounds change in a context-dependent manner due to coarticulation with neighboring sounds (e.g. Hillenbrand, Clark, & Nearey, 2001). A lexical-distributional learner hearing reliable differences between sounds in different words might erroneously assign coarticulatory variants of the same phoneme to different categories, having no other mechanism to deal with context-dependent variability. Such variability may need to be represented explicitly if an interactive learner is to categorize coarticulatory variants together.

Infants learn multiple levels of linguistic structure, and it is often implicitly assumed that these levels of structure are acquired sequentially. These simulations have instead investigated the optimal learning outcome in an interactive system using a nonparametric Bayesian framework that permits simultaneous learning at multiple levels. The results demonstrate that information from words can lead to more robust learning of phonetic categories, providing one example of how such interaction between domains might help make the problem of language acquisition more tractable.

Chapter 5

Lexical-distributional learning in human learners

5.1 Introduction

The computational modeling results from Chapter 4 demonstrate that for an ideal learner, taking word-sized units into account can provide useful information in phonetic category learning. A question remains as to whether human learners are sensitive to these cues. Preliminary evidence that infants are sensitive to this type of information comes from the experiments by Thiessen (2007), in which 15-month-old infants discriminated *taw* and *daw* in a word-learning task only when they were familiarized with additional object labels *tawgoo* and *dawbow*. These results are consistent with the interpretation that it was distinct lexical contexts that helped infants distinguish the sounds. However, Thiessen's study involved a word learning task, and the results may therefore have been influenced by the presence of potential referents. Because infants do not show clear evidence of mapping words to referents during the period when they are learning phonetic categories, it is important to show that word-level cues can affect phonetic categorization even in the absence of referential information.

In this chapter behavioral experiments test the lexical-distributional model's predictions in human learners. Experiment 1 tests learners' sensitivity to word-level information, and Experiment 2 tests whether this sensitivity extends to a situation in which the words are heard as part of a fluent stream of speech. Both experiments use minimally different syllables involving a vowel contrast (*tah* vs. *taw*). This focus on vowels complements the focus on stop consonants from Thiessen's experiment. Vowel categories typically have a higher degree of acoustic overlap than stop consonants, as is evident from a comparison of acoustic measurements of each type of category (e.g. Lisker & Abramson, 1964; Peterson & Barney, 1952). This suggests that inferences about vowel categories can stand to benefit most from supplementary word-level cues, because of the difficulty of recovering overlapping categories through purely distributional learning. Taken together with Thiessen's results, these experiments have the potential to provide converging evidence for a lexical-distributional phonetic category learning mechanism across different contrasts, tasks, and ages.

These experiments follow the design of the toy simulations from Section 4.3, in which the lexical-distributional model disambiguated the acoustically overlapping categories B and C when they appeared in distinct lexical contexts AB and DC, but did not disambiguate the categories when they appeared interchangeably in the same set of lexical contexts. In these experiments the syllables *gu*, *taw*, *tah*, and *li* take the place of categories A, B, C, and D, respectively. While these syllables do not differ along a single acoustic dimension, and thus do not directly correspond to the distributions associated with toy categories A, B, C, and D, they are consistent with the same general pattern in that *taw* and *tah* are acoustically similar, whereas *gu* and *li* are easily distinguishable based on distributional information.

In each experiment, half the participants hear a LEXICAL corpus containing either *gutah* and *litaw*, or *gutaw* and *litah*, but not both pairs of words. These participants hear *tah* and *taw* in distinct lexical contexts, with the specific pairings counterbalanced across participants. The other half hear a NON-LEXICAL corpus containing all four pseudowords. These participants hear *tah* and *taw* interchangeably in the same set of lexical contexts. The model predicts that participants who hear the LEXICAL corpus should separate the overlapping *tah* and *taw* categories because of their occurrence in distinct lexical contexts.

Although the lexical-distributional model makes a clear prediction that the LEXICAL corpus provides evidence for two separate categories, the information participants receive linking sounds and words is inherently ambiguous if one takes a broader view of phonetic category acquisition. This ambiguity arises because word contexts and phonological contexts are confounded in language input, so that participants can attribute acoustic differences between *tah* and *taw* to either lexical or phonological factors. Under a lexical interpretation, the pattern in the LEXICAL corpus arises because words in a language do not exhaust all possible phoneme sequences. Idiosyncratic differences across different lexical items occur because some lexical items contain one phoneme and some contain the other. For a learner that chooses this interpretation, consistent acoustic differences across lexical contexts provide a source of disambiguating evidence that the sounds belong to different categories.

Under a phonological interpretation, however, the pattern in the LEXICAL corpus can arise due to a process like vowel-to-vowel coarticulation or vowel harmony. These processes can cause the same vowel to be pronounced with different acoustic properties, where the acoustic realization depends on the preceding vowel. If participants hear the words *gutaw* and *litah*, it is possible to interpret *taw* and *tah* as representing the same underlying phoneme but being conditioned by the preceding phonological contexts /u/ and /i/. A learner hearing the LEXICAL corpus that interprets the acoustic differences as phonological should notice that the sounds are in complementary distribution, and should take the systematicity of the differences between *tah* and *taw* as evidence that the sounds are underlyingly identical. English-learning infants are sensitive to vowel harmony patterns at seven months, despite lack of exposure to these patterns, suggesting that the phonological interpretation of acoustic variability is available at the age when they are first acquiring phonetic categories (Mintz, Walker, Welday, & Kidd, in preparation).

Participants hearing the LEXICAL corpus do not have enough information to determine whether the acoustic differences between *tah* and *taw* depend on lexical or phonological contexts. They hear only one word with *tah* and one word with *taw*; this evidence is consistent with either possibility. This is representative of the situation in first language acquisition, because the confound between lexical and phonological interpretations of acoustic differences is present in the evidence that children receive when acquiring language. Upon hearing two similar sounds in different lexical contexts, language learners need to determine whether the acoustic differences arise from underlying phonemic differences or phonological processes. One way that learners might begin to distinguish between phonemic and allophonic differences is by the naturalness of the alternation. Many arguments have been made that "natural" phonological alternations are easier to learn than arbitrary alternations, though empirical investigations to date have yielded mixed results. Peperkamp et al. (2003) and Seidl and Buckley (2005) found that adults can learn both natural and unnatural phonological rules in an experimental task. However, Peperkamp, Skoruppa, and Dupoux (2006) showed higher levels of generalization when patterns of alternation involved natural classes, such as the voiced

stops [b], [d], and [g], than when they involved arbitrary groups of segments. Saffran and Thiessen (2003) showed a similar facilitation in infants for such patterns. Wilson (2003) demonstrated that natural phonological processes were generalized more easily than random processes to novel stimuli, and Wilson (2006) also found an effect of naturalness of a palatalization process in an artificial learning task with adults. The naturalness of an alternation therefore seems to have some effect in the ease with which certain patterns are learned. Sensitivity to the naturalness of the alternation can be examined in this experiment by analyzing differences between the two counterbalancing subconditions of the LEXICAL condition. In the *gutaw-litah* subcondition, the *gu* syllable with low F_2 and is paired with the *taw* syllable, which has lower F_2 than *tah*. Similarly, the *li* syllable with high F_2 is paired with the *tah* syllable with high F_2. This means the differences between *gu* and *li* are in the same direction as those between the *ta* syllables, making this a natural alternation. In the *gutah-litaw* subcondition, the pattern represents a less nataural alternation because the second formant in the *ta* syllable is shifted in the opposite direction from what would be predicted on the basis of the context syllable.

If participants interpret the acoustic differences between *tah* and *taw* as arising through lexical differences, as the lexical-distributional model would predict, then participants familiarized with a LEXICAL corpus should treat those sounds as different more often than participants familiarized with the NON-LEXICAL corpus. This pattern would represent the strongest evidence in support of the model, and would demonstrate that participants are sensitive to word-level cues that can help separate overlapping categories. If participants instead treat the acoustic differences as arising through vowel-to-vowel coarticulation or vowel harmony, then participants familiarized with a LEXICAL corpus should treat the sounds as different less often than participants familiarized with the NON-LEXICAL corpus. This result would be predicted to be stronger for participants familiarized with *gutaw-litah* than for participants familiarized with *gutah-litaw* due to the relative naturalness of the two types of alternations. Such a result would not provide direct evidence in support of the model, but would still indicate a general sensitivity to contextual information.

5.2 Experiment 1

5.2.1 Introduction

Experiment 1 tests whether adults are sensitive to word-level cues in a phonetic category learning task. The experiment is modeled after Maye and Gerken (2000), who tested adults' sensitivity to distributional information in a phonetic category learning task. In that experiment, participants were familiarized with isolated syllables whose stop consonants were drawn from either a unimodal or a bimodal distribution. During test, they were asked to make explicit judgments about whether the endpoint stimuli belonged to the same category in the language they just heard. Collecting explicit judgments ensured that the results reflected inferences about category membership rather than focusing only on low-level changes in discrimination. In that experiment, participants in the bimodal condition responded *different* significantly more often to pairs of endpoint stimuli than participants in the unimodal condition, indicating that they treated the stimuli as belonging to two categories.

Appearance in distinct lexical contexts is predicted to influence phonetic categorization in a way similar to hearing a bimodal distribution of sounds. Participants who hear stimuli in distinct lexical contexts should treat stimuli from the two categories as *different* more often than should participants who hear the sounds used interchangeably in the same set of lexical contexts.

5.2.2 Methods

Participants

Forty adult native English speakers with normal hearing from the Brown University community participated in this study. Participants were paid at a rate of $8/hour as compensation for their participation.

Stimuli

Stimuli consisted of an 8-point vowel continuum ranging from *tah* (/tɑ/) to *taw* (/tɔ/) and ten filler syllables: *bu*, *gu*, *ko*, *li*, *lo*, *mi*, *mu*, *nu*, *ro*, and *pi*. Several tokens of each of these syllables were recorded by a female native speaker of American English.

Tokens of *tah* and *taw* were analyzed for first and second steady state formant values and were found to differ systematically only by their second formant, F_2. An F_2 continuum was created based on formant values from these tokens, containing eight equally-spaced tokens along an ERB psychophysical scale (Glasberg & Moore, 1990). Steady state second formant values from this continuum are shown in Table 5.1. All tokens in the continuum had steady state values of F_1 = 818 Hz, F_3 = 2750 Hz, F_4 = 3500 Hz, and F_5 = 4500 Hz. For the first and third formants, these values were based on measurements from a recorded *taw* syllable. Bandwidths for the five formants were set to 130, 70, 160, 250, and 200, respectively, based on values given in Klatt (1980) for the /ɑ/ vowel.

To create tokens in the continuum, a source-filter separation was performed in Praat (Boersma, 2001) on a recorded *taw* syllable that had been resampled at 11000 Hz. The source was checked through careful listening and inspection of the spectrogram to ensure that no spectral cues remained to the original vowel. A 53.8 ms portion of aspiration was removed from the source token to improve its subjective naturalness as judged by the experimenter, resulting in a voice onset time of approximately 50ms.

Eight filters were created that contained formant transitions leading into steady-state portions. Formant values at the burst in the source token were F_1 = 750 Hz, F_2 = 1950 Hz, F_3 = 3000 Hz, F_4 = 3700 Hz, and F_5 = 4500 Hz. Formant transitions were constructed to move from these burst values to each of the steady-state values from Table 5.1 in ten equal 10 ms steps, then stay at steady-state values for the remainder of the token. These eight filters were applied to copies of the source file using the Matlab signal processing toolbox. The resulting vowels were then cross-spliced

Stimulus Number	Second Formant (Hz)
1	1517
2	1474
3	1432
4	1391
5	1351
6	1312
7	1274
8	1237

Table 5.1: Second formant values of stimuli used in Experiments 1-4.

with the unmanipulated burst from the original token. The stimuli were edited by hand to remove several clicks resulting from discontinuities in the waveform at formant transitions. These clicks were removed by splicing together zero-crossings on either end of the pitch period in which the discontinuity occurred, resulting in the removal of 17.90 ms total, encompassing four pitch periods, from three distinct regions in the formant transition portion of each stimulus. An identical set of regions was removed from each stimulus in the continuum. After splicing, the duration of each token in the continuum was 416.45 ms.

Four tokens of each of the remaining syllables were resampled at 11000 Hz to match the synthesized *ta* tokens, and the durations of these filler syllables were modified to match the duration of the *ta* tokens. The pitch of each token was set to a constant value of 220 Hz. RMS amplitude was normalized across tokens.

Bisyllabic pseudo-words *guta*, *lita*, *romu*, *pibu*, *komi*, and *nulo* were constructed through concatenation of these tokens. Thirty-two tokens each of *guta* and *lita* were constructed by combining the four tokens of *gu* or *li* with each of the eight stimuli in the continuum. Sixteen tokens of each of the four bisyllabic filler words were created using all possible combinations of the four tokens of each syllable.

Apparatus

Participants were seated at a computer and heard stimuli through Bose QuietComfort 2 noise cancelling headphones at a comfortable listening level.

Procedure

Participants were assigned to one of two conditions, the NON-LEXICAL condition or the LEXICAL condition, and completed two identical blocks. Each block contained a familiarization period followed by test.

During familiarization, each participant heard 128 pseudo-word tokens per block. Half of these consisted of one presentation of each of the 64 filler tokens (*romu*, *pibu*, *komi*, and *nulo*). The other half consisted of 64 experimental tokens (*guta* and *lita*). All participants heard each *ta* token from the continuum eight times per block, but the lexical contexts in which they heard these syllables differed across conditions. Participants in the NON-LEXICAL condition heard each *guta* and *lita* token once per block. Participants in the LEXICAL condition were divided into two subconditions. Participants in the *gutah-litaw* subcondition heard the 16 *guta* tokens made from steps 1-4 of the continuum and the 16 *lita* tokens made with steps 5-8 of the continuum twice per block. They did not hear *guta* tokens with steps 5-8 of the continuum or *lita* tokens made with steps 1-4. Conversely, participants in the *gutaw-litah* subcondition heard the 16 *guta* tokens with steps 5-8 of the continuum and the 16 *lita* tokens with steps 1-4 of the continuum twice per block, but did not hear *guta* tokens with steps 1-4 or *lita* tokens with steps 5-8. The order of presentation of these 128 pseudowords was randomized, and there was a 750 ms interstimulus interval between tokens.

During test, participants heard two syllables and were asked to make explicit judgments as to whether the syllables belonged to the same category in the language. The instructions were as follows:

> Now you will listen to pairs of syllables and decide which sounds are the same. For

example, in English, the syllables CAP and GAP have different sounds. If you hear two different syllables (e.g. CAP-GAP), you should answer DIFFERENT, because the syllables contain different sounds. If you hear two similar syllables (e.g. GAP-GAP), you should answer SAME, even if the two pronunciations of GAP are slightly different.

The syllables you hear will not be in English. They will be in the language you just heard. You should answer based on which sounds you think are the same in that language. Even if you're not sure, make a guess based on the words you heard before.

Participants were then asked to press buttons corresponding to *same* or *different* and to respond as quickly and accurately as possible.

The test phase examined three contrasts: *ta1* vs. *ta8* (far contrast), *ta3* vs. *ta6* (near contrast), and *mi* vs. *mu* (control). Half the trials were *different* trials containing one token of each stimulus type in the pair, and the other half were *same* trials containing two tokens of the same stimulus type. Each trial had a 750 ms interstimulus interval between the two syllables. For *same* trials involving *ta* stimuli, the two stimuli were simply the same token repeated twice. For *same* trials involving *mi* and *mu*, the two stimuli were non-identical tokens of the same syllable, to ensure that participants were correctly following the instructions to make explicit category judgments rather than lower-level acoustic judgments. Participants heard 16 *different* and 16 *same* trials for each *ta* contrast (far and near) and 32 *different* and 32 *same* trials for the control contrast, per block. Responses and reaction times were recorded for each trial.

5.2.3 Results

Responses were excluded from the analysis if the reaction time was more than two standard deviations from a participant's mean reaction time for a particular response on a particular class of trial in a particular block. The sensitivity measure d' (Green & Swets, 1966) was computed from the

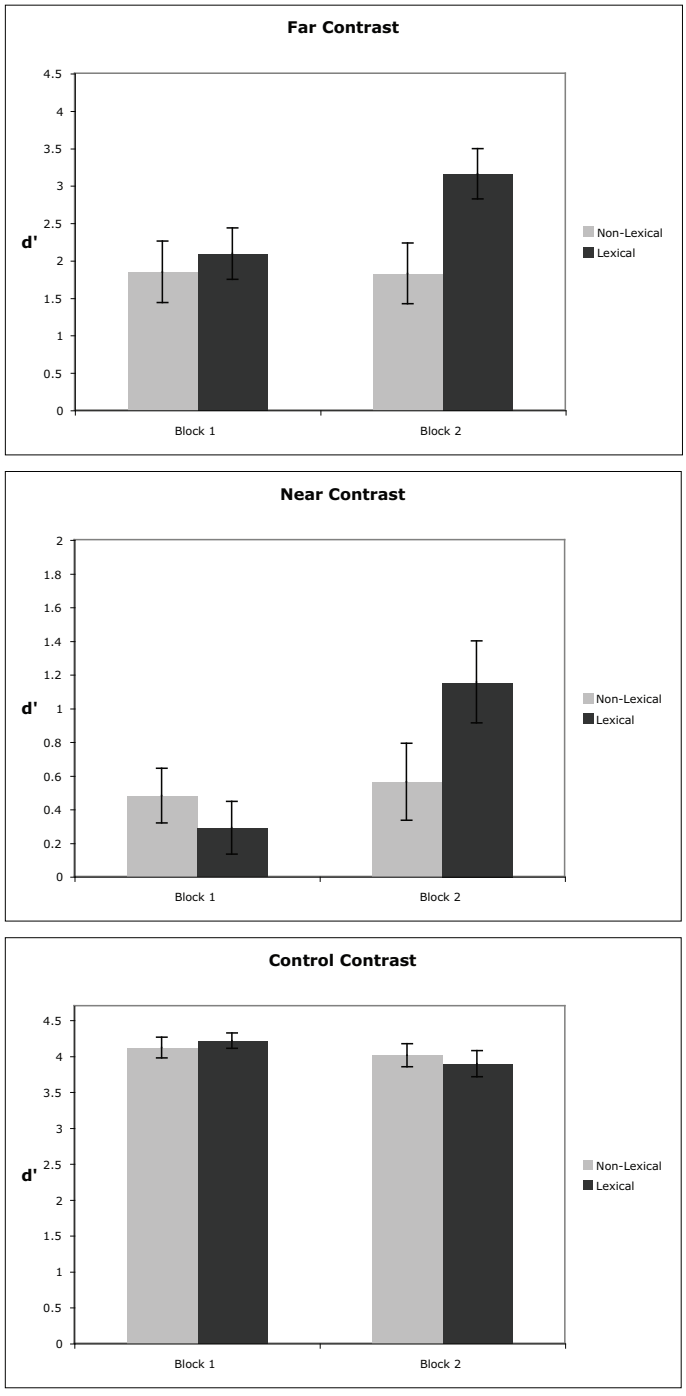

Figure 5.1: Adults' sensitivity to category differences in the (a) far contrast, (b) near contrast, and (c) control trials in Experiment 1.

remaining responses for each contrast in each block. A value of 0.99 was substituted for any trial type in which a participant responded *different* on all trials, and a value of 0.01 was substituted for any trial type in which a participant responded *same* on all trials.

A 2×2 (condition × block) mixed ANOVA was conducted for each contrast. For the far contrast and the near contrast, the analysis yielded a main effect of block ($F(1, 38) = 10.42$, $p = 0.003$, far contrast; $F(1, 38) = 17.99$, $p < 0.001$, near contrast) and a significant interaction ($F(1, 38) = 11.25$, $p = 0.002$, far contrast; $F(1, 38) = 12.30$, $p = 0.001$, near contrast). This interaction reflected the larger increase in d' scores of participants in the LEXICAL condition than participants in the NON-LEXICAL condition. There was no main effect of condition for either contrast. Tests of simple effects showed no significant effect of condition in the first block; there was a significant effect of condition in the second block for the far contrast ($t(38) = 2.54$, $p = 0.03$, Bonferroni corrected), but this comparison did not reach significance for the near contrast. A 2×2×2 (condition × block × contrast) ANOVA[1] confirmed that the near and far contrasts patterned similarly, showing main effects of block ($F(1, 38) = 20.65$, $p < 0.001$) and contrast ($F(1, 38) = 62.84$, $p < 0.001$) and a block×condition interaction ($F(1, 38) = 18.18$, $p < 0.001$) but no interactions involving contrast.

On control trials, the analysis yielded a main effect of block ($F(1, 38) = 5.90$, $p = 0.02$), reflecting the fact that d' scores were reliably lower in the second block. This decrease in d' scores between blocks was in the opposite direction from the increase in d' scores between blocks on experimental trials. There was no difference between groups and no significant interaction. Sensitivity to category differences was high, with an average d' measure of 4.17 on the first block and 3.95 on the second block, indicating that participants were performing the correct task.

Analyses showed no significant differences between the *gutah-litaw* and *gutaw-litah* subconditions, and no interactions between subcondition and block, for any of the three contrasts.

[1] The control contrast cannot be included in this direct comparison because the tokens in the *same* control trials were acoustically different, whereas the tokens in *same* experimental trials were acoustically identical.

5.2.4 Discussion

This experiment was designed to test whether adults are sensitive to word-level cues to phonetic category membership in an artificial language learning task. The lexical-distributional model predicts that participants in the LEXICAL group should be more likely to treat the *tah* and *taw* stimuli as belonging to different categories, whereas participants in the NON-LEXICAL group should be more likely to treat these stimuli as belonging to the same category. This predicted pattern was obtained after the second block of training: Participants in the LEXICAL condition showed higher d' scores than participants in the NON-LEXICAL condition. These results support the lexical-distributional model, showing that adults interpret acoustic variability differently on the basis of word-level information and that appearance in distinct lexical contexts leads to relatively higher sensitivity to category differences.

The two groups' indistinguishable performance after the first training block provides strong evidence that these differences in sensitivity were the result of learning over the course of the experiment. The direction of learning, however, cannot be inferred from these data. One possibility is that the LEXICAL group learned over the course of the experiment to treat the experimental stimuli as different. Under this interpretation, their increase in d' scores in the second block would reflect category learning that resulted from the specific word-level information they received about these sounds during familiarization. Another possibility is that the increase in d' scores in the LEXICAL group reflected perceptual learning that arose through simple exposure to the sounds, and that this perceptual learning was not apparent in the NON-LEXICAL group because those participants learned to treat the experimental stimuli as the same based on their interchangeability in words. These two possibilities cannot be distinguished without a baseline measure of categorization that is independent of familiarization condition. Thus, it is not known from this experiment whether the lexical information serves to separate overlapping categories or merge distinct acoustic categories that are used interchangeably, but either possibility is consistent with the hypothesis that

participants use word-level information to constrain their interpretation of phonetic variability.

Participants in Experiment 1 showed evidence of interpreting sounds as different based on their appearance in distinct lexical contexts. The patterns obtained here resemble the results from Thiessen (2007), but show that referents are not required for interactive learning. Taken together with previous results showing sensitivity to distributional cues (Maye & Gerken, 2000; Maye et al., 2002), these results indicate that human learners attend to the types of cues that are necessary to achieve lexical-distributional learning

5.3 Experiment 2

5.3.1 Introduction

Experiment 1 showed that adults can use information from words when reasoning about category membership of similar sounds. This suggests a role for word-level information in phonetic category acquisition. However, in Experiment 1, words were presented in isolation during the familiarization phase. This contrasts with the input data to real-world language learners, in which words are embedded in fluent speech. Experiment 2 tests whether learners are sensitive to word-level cues to phonetic categorization when the words are heard as part of a fluent stream of speech.

While some recent studies of word segmentation have used tokens that vary acoustically (Pelucchi, Hay, & Saffran, 2009b, 2009a), little is known about how word segmentation processes interact with acoustic variability. An interactive account would predict that the two processes proceed in parallel, each influencing the learning outcome in the other domain. Phonetic information seems to influence word segmentation, as indicated by infants' sensitivity to allophonic cues in word segmentation tasks (Jusczyk, Hohne, & Bauman, 1999). However, it is not known whether the words feed back to affect phonetic categorization.

Emberson, Liu, and Zevin (2009) tested whether reliable transitional probabilities among categories of sounds would help learners assign sounds to their appropriate categories. Their stimuli

involved six different tokens of each of four categories of non-speech stimuli, which they labeled E1, E2, H1, and H2. Each pair of categories (E1-E2 and H1-H2) was potentially confusible. Preliminary multidimensional scaling analyses indicated that before training, categories E1 and E2 had substantial separation in perceptual space, whereas categories H1 and H2 were heavily overlapping in perceptual space. To see whether transitional probabilities affected category learning, participants were trained with a pattern in which the four categories were organized into two two-syllable words, which were presented in a random sequence in a fluent stream of speech. Each sound category contained six different sound tokens. Transitional probabilities were defined in terms of categories, such that transitional probabilities could give cues to which sounds belonged to which categories. However, participants did not show evidence of updating their beliefs about category membership in response to these transitional probability cues. Across three training scenarios in which the "words" were either E1E2 and H1H2, E1H1 and E2H2, or E1H1 and H2E2, patterns of discrimination between words and non-words supported a model in which listeners had been tracking transitional probabilities for E1 and E2 separately, but had treated H1 and H2 as a single category. The participants did not show any evidence of using the slight differences in acoustics between H1 and H2, combined with information from lexical contexts, to determine that those were different categories.

Learners' failure to separate the H1 and H2 categories on the basis of transitional probabilities appears to contradict the lexical-distributional model's predictions. However, it is possible that this failure resulted from specific aspects of their experiment, such as the use of non-speech stimuli or differences in task difficulty. In order to succeed at their task, participants need to use word-level information to update beliefs about sound category membership, and then use beliefs about sound categories to update beliefs about word boundaries. The phonetic categorization task from Experiment 1 allows participants to succeed by performing only the first of these inferences, making it an easier task.

Experiment 2 tests participants' ability to use word-level information to constrain phonetic category learning when words are presented as part of a fluent stream of speech. This experiment uses

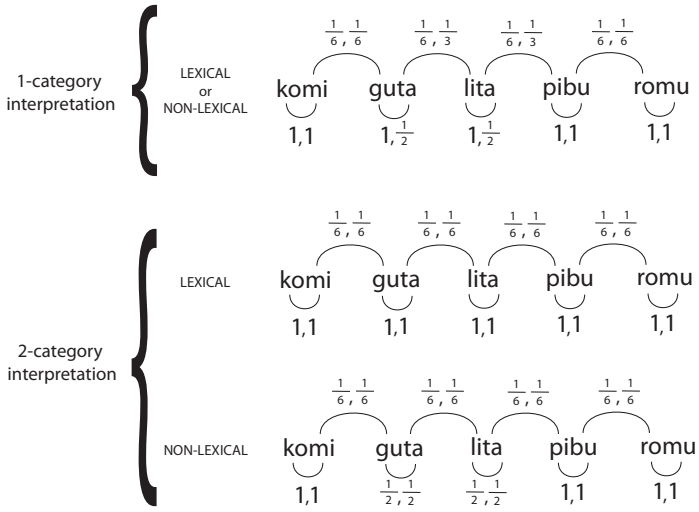

Figure 5.2: Transitional probabilities under each interpretation of category membership. For each transition, the first number gives the forward transitional probability and the second number gives the backward transitional probability.

the stimuli from Experiment 1, which were shown to facilitate lexical-distributional learning when presented in isolation. The experiment therefore tests whether the results found previously can be generalized to a more complex language learning situation.

For this experiment, word-level cues to phonetic category membership were embedded in a fluent stream of speech. The fluent stream of speech was created through concatenation of word tokens such that transitions between words were always more likely than transitions within words. It is well established that learners can use transitional probabilities for word segmentation (e.g. Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996; Aslin et al., 1998), and replicating this finding is not central to the current experiment. Instead, the aim is to see whether learners use the words they segment from fluent speech to constrain their phonetic categorization. Thus, no attempt was made in this experiment to avoid having the same word twice in a row, or to mask the beginning or end of the stream with extra syllables; instead, the transitional probability cues were made as strong as possible within the constraints of the stimulus set.

Specific transitional probabilities in each corpus differed depending on whether the *tah* and *taw* syllables were interpreted as one or two categories. In addition, the transitional probabilities under the two-category interpretation were different between the LEXICAL and NON-LEXICAL corpora. Probabilities for each of the possible interpretations are shown in Figure 5.2. Each set of forward and backward transitional probabilities is expected to contain enough information to support segmentation. Although the cues to word boundaries present in backward transitional probabilities are as weak as 0.33 versus 0.5 in some cases, this difference has previously been sufficient in facilitating discrimination between words and part-words in a segmentation task (Endress & Mehler, 2009). Forward transitional probability cues are stronger than this in all cases. Adult and infant learners can track both forward and backward transitional probabilities (Perruchet & Desaulty, 2008; Pelucchi et al., 2009b, 2009a), and the forward and backward transitional probability cues together are thus expected to be sufficient to allow learners to segment the speech into component words.

It is not necessary to segment words successfully in order to notice a difference between the LEXICAL and NON-LEXICAL conditions. Because in the LEXICAL condition the syllables *gu* and *li* are predictive of the syllables *tah* and *taw*, simply attending to transitional probabilities should allow performance to differ between the two conditions. As argued earlier in this chapter, however, transitional probabilities can be cues to either coarticulatory or phonemic differences, so that the strongest evidence for categorizing stimuli into different categories should come from performing a word-level analysis.

5.3.2 Methods

Participants

Forty adult native English speakers with normal hearing from the Brown University community participated in this study. Participants were paid at a rate of $8/hour as compensation for their participation.

Stimuli and apparatus

Stimuli and apparatus were identical to those used in Experiment 1.

Procedure

Participants were assigned to the NON-LEXICAL condition or the LEXICAL condition and completed two identical blocks. Each block contained a familiarization period followed by test. After the second block participants completed a segmentation post-test.

During familiarization, each participant heard a stream of syllables with no pauses between them. The stream consisted of 64 experimental tokens (32 each of *guta* and *lita*) and 128 filler tokens (32 each of *romu*, *pibu*, *komi*, and *nulo*) in random order. Participants were simply told to listen to the speech, and were not told that it contained words.

As in Experiment 1, all participants heard each *ta* token from the continuum eight times per block, but the lexical contexts in which they heard these syllables differed across conditions. Participants in the NON-LEXICAL condition heard each token of *guta* and *lita* once per block. Participants in the LEXICAL condition were divided into two subconditions. Participants in the *gutah-litaw* subcondition heard the 16 *guta* tokens made from steps 1-4 of the continuum and the 16 *lita* tokens made with steps 5-8 of the continuum twice per block. Participants in the *gutaw-litah* subcondition heard the 16 *guta* tokens with steps 5-8 of the continuum and the 16 *lita* tokens with steps 1-4 of the continuum twice per block.

The test phase was identical to the test phase in Experiment 1.

During post-test, participants heard pairs of two-syllable words and were asked which was a word in the language they just heard. Each trial contained one of the six words and one of six part-words (*taro*, *muko*, *minu*, *logu*, *tapi*, and *buli*). Each of the 36 pairings between a word and a part-word was presented twice, counterbalanced for presentation order, for a total of 72 post-test trials.

5.3.3 Results

Participants in both conditions segmented words significantly above chance levels overall, with an average of 43 of 72 correct in the NON-LEXICAL condition ($t(19) = 10.96$, $p < 0.001$) and 46 of 72 correct in the LEXICAL condition ($t(19) = 17.05$, $p < 0.001$). However, there was considerable variability in segmentation performance, with many individual scores indistinguishable from chance level. Because the focus of this experiment was to examine the influence of segmented words on phonetic categorization, separate analyses were conducted for those participants that did and did not segment successfully. The first analysis included only participants whose overall segmentation scores were significantly above chance in a one-tailed binomial test (44 or above). This included 10 participants in the non-lexical condition and 11 participants in the lexical condition. A separate analysis was conducted on the 10 participants in the non-lexical condition and 9 participants in the lexical condition whose segmentation scores were indistinguishable from chance level.

Same-different responses were excluded from the analysis if the reaction time was more than two standard deviations from a participant's mean reaction time for a particular response on a particular class of trial in a particular block. The sensitivity measure d' was computed from the remaining responses for each contrast in each block. A value of 0.99 was substituted for any trial type in which a participant responded *different* on all trials, and a value of 0.01 was substituted for any trial type in which a participant responded *same* on all trials.

Separate 2×2 (condition × block) mixed ANOVAs were conducted for each contrast. For participants who segmented significantly above chance, analysis of the far contrast yielded a significant main effect of condition ($F(1, 19) = 4.76$, $p = 0.04$), with participants in the non-lexical condition showing higher d' scores than participants in the lexical condition, but no main effect of block and no interaction. This effect of condition was in the opposite direction of that found in the second block of Experiment 1. For the near contrast the effect of condition was also significant ($F(1, 19) = 5.34$, $p = 0.03$), with participants in the non-lexical condition again exhibiting higher d' scores. There

was no main effect of block, but the interaction between these two factors was marginally significant ($F(1, 19) = 3.04$, $p = 0.10$). A 2×2×2 (condition × block × contrast) mixed ANOVA was conducted to determine whether the far and near contrasts patterned similarly. This analysis yielded significant main effects of contrast ($F(1, 19) = 27.77$, $p < 0.001$) and condition ($F(1, 19) = 5.49$, $p = 0.03$) but no other main effects or interactions, indicating parallel results for both contrasts.

For the control contrast, there was a marginally significant effect of block ($F(1, 19) = 3.98$, $p = 0.06$), which reflected a similar decrease in d' scores to that found in Experiment 1. There was no effect of condition and no interaction.

Next, potential differences between the *gutah-litaw* and *gutaw-litah* subconditions were examined. Analysis of the far contrast revealed a marginally significant main effect of block ($F(1, 9) = 3.80$, $p = 0.08$) and a significant block×subcondition interaction ($F(1, 9) = 7.61$, $p = 0.02$). Analysis of the near contrast yielded a main effect of block ($F(1, 9) = 6.37$, $p = 0.03$) and a block×subcondition interaction ($F(1, 9) = 8.41$, $p = 0.02$). Both interactions reflected a marginally significant increase in d' scores in the *gutah-litaw* group between the first and second blocks ($t(5) = 3.04$, $p = 0.06$ for the far contrast; $t(5) = 3.04$, $p = 0.06$ for the near contrast; Bonferroni corrected), paired with a slight non-significant decrease in d' scores in the *gutaw-litah* group. A 2×2×2 (subcondition x block x contrast) ANOVA confirmed these effects, with a main effect of contrast ($F(1, 9) = 6.29$, $p = 0.03$), a marginally significant main effect of block ($F(1, 9) = 5.08$, $p = 0.05$), and a significant block×subcondition interaction ($F(1, 9) = 8.98$, $p = 0.02$). The three-way interaction was marginally significant ($F(1, 9) = 4.48$, $p = 0.06$), suggesting that the size of the block×subcondition interaction may have been larger for the far contrast than for the near contrast, even though the interactions for both contrasts were significant and in the same direction. Analysis of the control contrast yielded no significant effects, indicating indistinguishable patterns of performance by participants in the two subconditions.

Overall, these analyses revealed that for participants who segmented significantly above chance

level, the NON-LEXICAL group showed greater sensitivity to category differences than the LEXICAL group, and the lack of block×condition interaction suggests that this remained constant over the course of the experiment. The two subgroups of the LEXICAL group showed different behavior, however, as indicated by the significant block×subcondition interaction. Sensitivity to category differences increased in the *gutah-litaw* subcondition more than in the *gutaw-litah* subcondition between blocks.

A second set of analyses examined those participants whose segmentation scores were not significantly different from chance. For the far contrast, these participants showed a main effect of block ($F(1, 17) = 4.99$, $p = 0.04$), reflecting higher d' scores in the second block, but no effect of condition and no interaction. For the near contrast the analysis revealed no significant or marginally significant effects. A 2×2×2 (condition × block × contrast) mixed ANOVA yielded significant main effects of contrast ($F(1, 19) = 34.20$, $p < 0.001$) and block ($F(1, 19) = 5.24$, $p = 0.04$). Although the three-way interaction was marginally significant ($F(1, 19) = 3.42$, $p = 0.08$), reflecting a larger increase in d' scores in the NON-LEXICAL group on the far contrast but in the LEXICAL group on the near contrast, this analysis suggests that participants patterned similarly for the two contrasts.

Analysis of the control contrast revealed a significant effect of block ($F(1, 19) = 8.08$, $p = 0.01$), reflecting lower d' scores in the second block; this was in the opposite direction of the effect of block found for the far contrast but mirrors the decline in d' scores found for the control contrast in Experiment 1 and for the participants segmenting above chance in Experiment 2.

There were no significant differences between the *gutah-litaw* and *gutaw-litah* subconditions on the far contrast or the near contrast, suggesting that the two subconditions of the LEXICAL group patterned similarly among those participants whose segmentation scores were at chance level.

5.3.4 Discussion

The effect obtained in this experiment is in the opposite direction of that predicted, with participants in the NON-LEXICAL group exhibiting higher d' scores than participants in the LEXICAL group

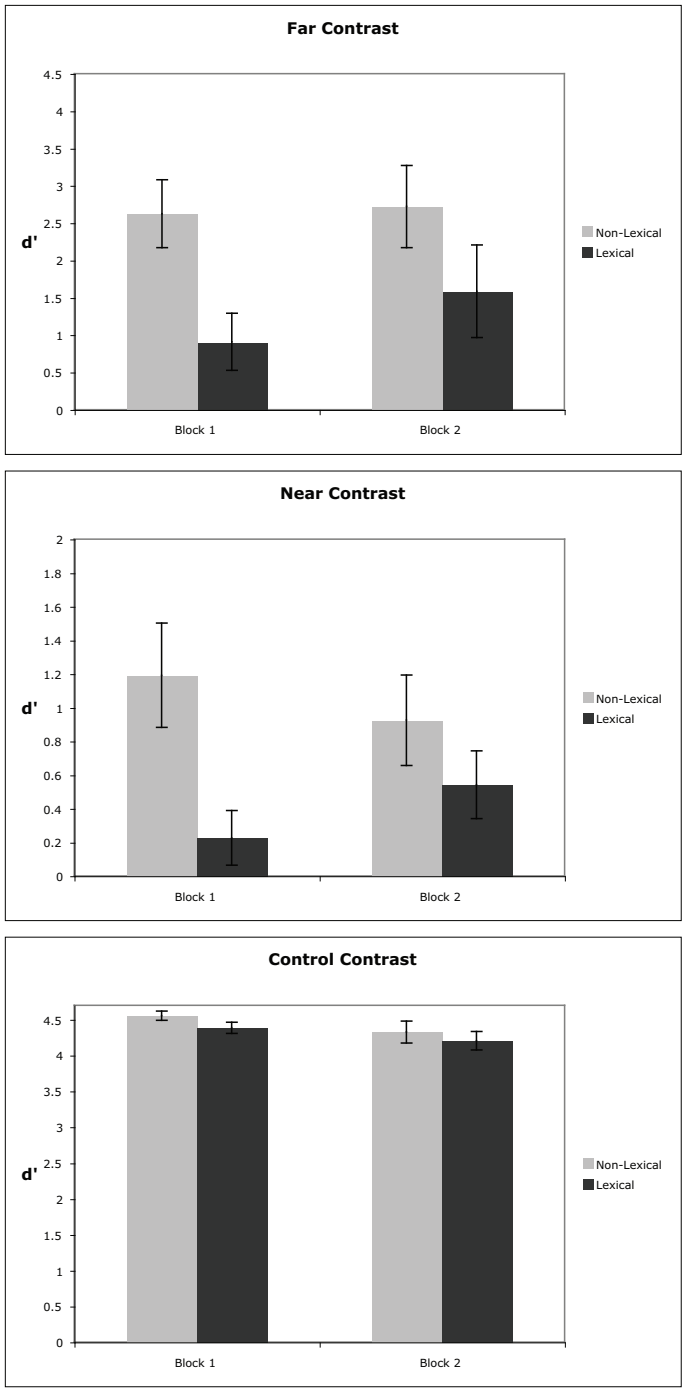

Figure 5.3: Sensitivity to category differences in the (a) far contrast, (b) near contrast, and (c) control trials for participants whose segmentation performance was significantly above chance in Experiment 2.

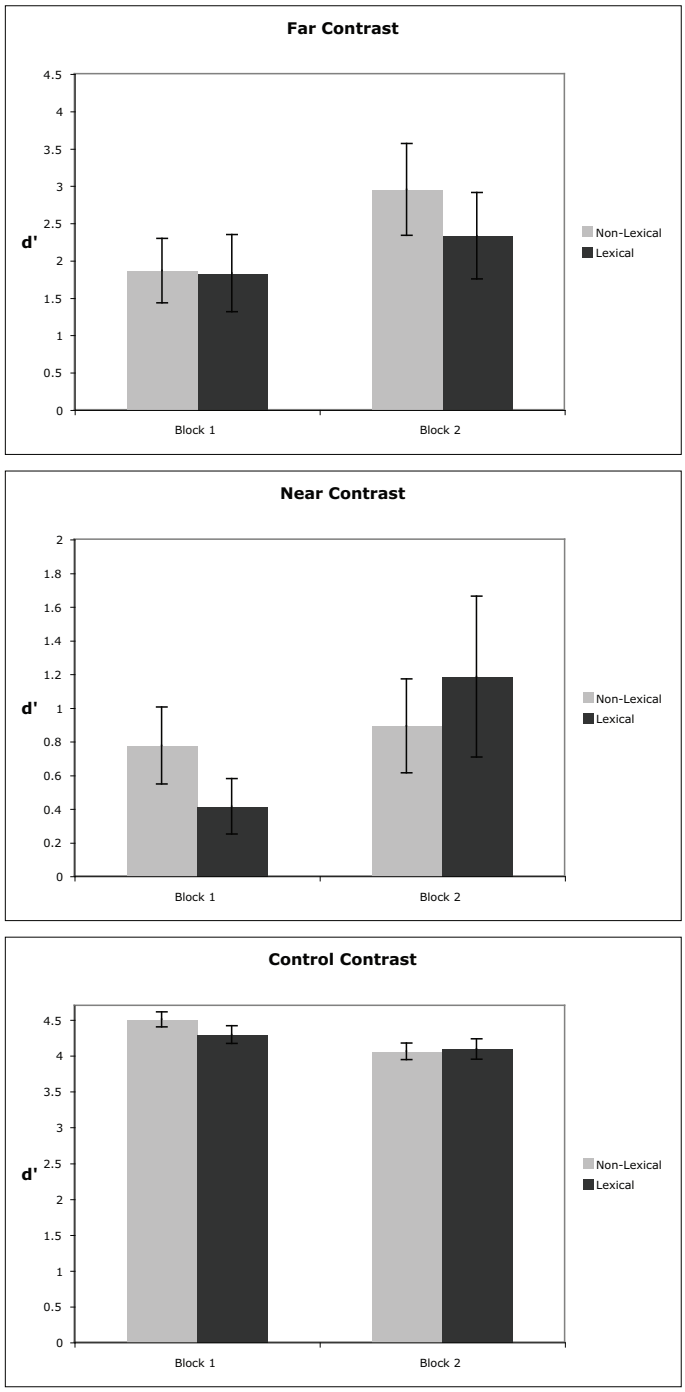

Figure 5.4: Sensitivity to category differences in the (a) far contrast, (b) near contrast, and (c) control trials for participants whose segmentation performance was not significantly above chance in Experiment 2.

for both the far and near contrasts. These results contrast sharply with those of Experiment 1. In Experiment 1, participants used the distribution of vowels in distinct lexical contexts to separate overlapping categories, labeling *tah* and *taw* as different more often by the end of the experiment when they occurred in different words. In Experiment 2, participants hearing the same two words embedded in fluent speech seemed to use that same distribution as evidence that *tah* and *taw* were part of the same category. This occurred despite the fact that these participants were above chance at discriminating words from part-words.

To explore the reversal, a post hoc analysis was conducted to evaluate segmentation performance of just those post-test trials in which participants had to choose between one of the target words, *guta* or *lita*, and a part-word. All participants were included in this analysis. Results showed that although participants segmented above chance overall, their preference for words over part-words in those trials that contained the critical items *guta* and *lita* was not statistically reliable, averaging 13 of 24 correct in both the LEXICAL and NON-LEXICAL conditions. Thus, despite segmenting above chance overall, participants in the LEXICAL condition may not have extracted the relevant words reliably enough to use them when learning about the phonetic categories of this language.

Nevertheless, the results from those participants who segmented above chance level overall do show sensitivity to higher-level structure. This is evidenced by the significant differences between the LEXICAL and NON-LEXICAL conditions, which can only have arisen through sensitivity to the contexts in which the *ta* sounds appeared. The results suggest that the participants in the LEXICAL condition from Experiment 2 who segmented successfully adopted a phonological interpretation of acoustic differences, assigning sounds to the same category more often if the differences between those sounds were predictable based on phonological context. Evidence that this behavior is linked to phonetic naturalness comes from the block×subcondition interaction. Participants in the *gutaw-litah* subcondition, who heard a phonologically natural pattern, continued to interpret the *taw* and *tah* syllables as belonging to a single category in the second block, whereas participants in the *gutah-litaw* subcondition showed a marginally significant increase in d' scores, to bring their scores closer

to those of participants in the NON-LEXICAL condition.

The analysis of participants who failed to segment significantly above chance revealed no difference between conditions. This contrasted with the analysis of participants who segmented above chance, and suggests that the difference between conditions in this experiment arose through a mechanism related to word segmentation. Given that participants did not show evidence of segmenting those words containing *ta* syllables, this relationship is unlikely to have been direct, in that the segmented words provided evidence for phonetic categorization. Instead, the relationship between segmentation and assignment of the *ta* syllables to a single category may have been indirect, in that participants who attended to transitional probabilities for word segmentation also interpreted those same transitional probabilities as evidence for contextually conditioned variation.

If the participants in Experiment 2 effectively had access only to transitional probability information, not to word-level information, the results of Experiment 2 provide evidence that participants can use transitional probability cues to identify sounds that occur in complementary distribution and can assign the sounds to a single phonological category on the basis of those distributional cues. The participants that did not segment words successfully did not show evidence of attending to these transitional probability cues, either for word segmentation or for identifying elements in complementary distribution. These results therefore strengthen the conclusions drawn by White et al. (2008) that transitional probabilities are an important cue related to identifying patterns of complementary distribution.

Note that these results contrast with those from Emberson et al. (2009), who argued that participants do not update their phonetic categorization when hearing transitional probability cues to words. The participants in Experiment 2 did modify their phonetic categorization of the relevant sounds, labeling those sounds as *same* more often when the sounds occurred in different lexical contexts. The results of this experiment provide evidence for the influence of top-down information, even if they do not support the qualitative predictions of the lexical-distributional model.

5.4 General discussion

The lexical-distributional model predicts that learners use distinct lexical contexts to separate neighboring phonetic categories, treating sounds as different when they occur consistently in different words. The results of Experiment 1 provide evidence for this type of learning: participants familiarized with the LEXICAL corpus showed evidence of learning over the course of the experiment to treat the *tah* and *taw* categories as different. This indicates sensitivity to word-level information for disambiguating acoustically similar phonetic categories. In this experiment, no evidence was found that learners were treating the acoustic differences as arising through phonological or coarticulatory processes, even though the information contained in the corpus was potentially ambiguous.

Learners hearing words embedded in fluent speech showed a qualitatively different pattern in Experiment 2, treating sounds as belonging to the same category more often when they occurred consistently in different contexts. Here there was evidence of an effect of phonetic naturalness, suggesting that participants may have relied on a phonological interpretation of acoustic variability. Participants in the *gutah-litaw* subcondition showed a marginally significant increase in d' scores across the two blocks when hearing a phonologically unnatural alternation, whereas participants in the *gutaw-litah* subcondition did not increase their sensitivity to category differences when hearing a phonologically natural alternation. There was no evidence in this experiment that participants separated phonetic categories that occurred in distinct lexical contexts. However, this may have been because the participants in this experiment did not learn the relevant words, performing at chance on those items during the segmentation post-test. Thus, participants may not have been receiving the word-level information that is necessary for lexical-distributional learning.

Although segmentation levels for *guta* and *lita* were not significantly above chance in Experiment 2, they were numerically above chance. This raises the possibility that participants in Experiment 2 were behaving like those in Experiment 1, but that they received less word-level evidence because this evidence only began accumulating after they have solved the segmentation problem. In

support of this, note that participants in the LEXICAL condition from Experiment 1 increased their d' scores from the first to the second block, and a similar marginally significant trend appeared for participants in the *gutah-litaw* subcondition from Experiment 2. It is possible that had participants from Experiment 1 been tested partway through the first block, they might have shown a pattern similar to participants from Experiment 2. Conversely, participants from Experiment 2 might have shown a pattern similar to Experiment 1 after more familiarization than just two blocks. The precise trajectory that would be predicted by this type of explanation, however, remains unexplained. In particular, this would indicate that participants first decrease, then increase, their beliefs that *tah* and *taw* represent different categories upon hearing those sounds in distinct lexical contexts. It is possible that this sort of trajectory could represent a shift from a phonological explanation to a word-based explanation of the acoustic differences between *tah* and *taw*. However, note that only participants in the *gutah-litaw* subcondition, but not the *gutaw-litah* subcondition, increased their sensitivity to category differences between the two blocks. This contrasts with the indistinguishable behavior of participants in these two subconditions in Experiment 1 and suggests that participants' interpretation of these patterns in Experiment 2 is qualitatively different from their behavior in Experiment 1.

Furthermore, even when participants in an experiment are able to segment the relevant test items above chance, it is not clear that they have access to word-level information involving those items. Endress and Mehler (2009) familiarized adult participants with a fluent stream of speech in which the only cues to word boundaries were transitional probabilities. The corpus was composed of six trisyllabic words whose forms were carefully constructed such that there existed "phantom words", which never appeared in the corpus, but which had the same transitional probabilities between successive syllables as the words actually appearing in the corpus. Results showed that although participants successfully discriminated words from part-words, they were at chance at distinguishing words from phantom words. This suggested that participants had retained information about transitional probabilities between syllables but had not retained any further information about

the words cued by those transitions. The information gathered from hearing a stream of speech in which words are cued by transitional probabilities may be qualitatively different from the information that can be gathered from hearing isolated words.

As argued by Endress and Mehler (2009), however, this failure to store word forms that are cued by transitional probabilities does not mean that such forms are unavailable to infants learning language. Participants in their experiment succeeded at discriminating words from phantom words when given word-final lengthening cues in addition to transitional probability cues, suggesting that this failure to store word forms is specific to settings in which transitional probabilities provide the only cues to word boundaries. Young infants can use a variety of cues, such as stress and allophonic cues, to segment words from fluent speech (Jusczyk, Houston, & Newsome, 1999; Jusczyk, Hohne, & Bauman, 1999). The limitation found in an experimental setting, when transitional probabilities are the only cues to word boundaries, would not necessarily apply to their more naturalistic language learning environment.

Word-level cues to language learners are inherently ambiguous. In the stimuli used in these experiments, acoustic differences between the *tah* and *taw* syllables might have arisen either through random acoustic variability, systematic phonemic differences between words, or systematic phonetic and phonological processes such as vowel harmony or vowel-to-vowel coarticulation. The NON-LEXICAL corpus provided evidence for the first of these possibilities, whereas the LEXICAL corpus provided evidence that was consistent with either of the latter two possibilities. Participants showed evidence of being sensitive to both of these possible interpretations. When hearing isolated words, participants in the LEXICAL condition treated the variability as arising through phonemic differences. However, when hearing words embedded in fluent speech, participants treated the same variability as occurring within a single phonemic category. The precise conditions under which human learners adopt a word-level or phonological interpretation of acoustic differences remains an important question for future investigation.

Chapter 6

Conclusions

6.1 Summary

Computational and empirical investigations tested the hypothesis that the words learners segment from fluent speech can provide useful information to guide phonetic category acquisition. In Chapter 4, a hierarchical Bayesian model was introduced that could learn multiple layers of structure simultaneously, acquiring a phonetic category inventory as well as categorizing word tokens into lexical items. Simulations indicated that when the structure of the lexicon matched the learner's assumptions about the lexicon, interactive learning drastically improved phonetic and lexical categorization over baseline models in each domain. When tested on words from the English lexicon, the interactive lexical-distributional model again showed improved phonetic categorization performance and, with some parameter combinations, also showed improved lexical categorization performance. However, the model showed errors as a result of the specific characteristics of the English corpus data, merging sets of lexical items and hypothesizing extremely variable phonetic categories to account for the high acoustic variability in those merged lexical items. Overall, these simulations showed that lexical-distributional learning can in principle lead to a better learning outcome but suggest the importance of other aspects of lexical structure, such as phonotactics and semantic information, that may also guide infants' learning (Jusczyk et al., 1994; Yeung & Werker, 2009).

Experiments in Chapter 5 examined to what extent real learners resembled the ideal lexical-distributional learner in their use of word-level information during phonetic category learning. Experiment 1 showed that adults behave qualitatively like lexical-distributional learners, using word-level information to separate similar sounds into distinct phonetic categories, even in the absence of referential information. These findings suggest that word-level cues are available to language learners, enabling learners to use such cues during phonetic category acquisition. Experiment 2 tested adults' use of word-level cues to phonetic category membership in a more complex language learning situation that required word segmentation as well as phonetic category assignment. Here behavior was reversed from that found in Experiment 1: participants assigned sounds to different

categories more often when they heard those sounds interchangeably in the same lexical contexts. This behavior only emerged for those participants who performed significantly above chance levels in word segmentation. One potential explanation for this reversal is that participants either did not segment the relevant words, or did not treat the segmented words as units (cf. Endress & Mehler, 2009). Transitional probabilities are useful cues to phonological structure that seem to be exploited by language learners (Peperkamp, Le Calvez, et al., 2006; White et al., 2008), so attending only to transitional probabilities may have led learners to favor a phonological interpretation over a word-level interpretation. The results of this experiment, when compared to the results of Experiment 1, underscore the complexity of speech sound category learning in a system in which multiple factors, such as phonemic and contextual differences, can influence a sound's acoustics.

6.2 Model extensions

6.2.1 Model assumptions

The lexical-distributional model was built to illustrate how feedback from a developing word-form lexicon can improve the robustness of phonetic category acquisition. The hierarchical nonparametric Bayesian framework was chosen for implementing this interactive model because it allows simultaneously learning of multiple layers of structure, with information from each layer affecting learning in the other layer in a principled way.

Whereas the idea of interactive learning is central to the work presented here, other aspects of the model represented simplifications that were made for computational simplicity. For example, speech sounds were represented as a pair of static formant values, rather than a set of acoustic values that vary over time. This is the same assumption that was made previously in models of phonetic category acquisition (McMurray et al., 2009; Vallabha et al., 2007), but is a gross oversimplification from the real speech that infants encounter. Infants actually have two phonetic categorization problems: the problem of learning boundaries in acoustic space, which was examined here, and a second problem of

learning temporal boundaries between phones. The simplified representation used in this and other models allows investigation of the category learning problem in acoustic space without requiring a model of temporal segmentation.

Still other decisions were made arbitrarily, such as the use of phones, rather than syllables, as the basic unit of the model. In the quantitative framework from Chapter 4, the focus on individual phones allowed these basic acoustic units to be represented in just two dimensions, corresponding to the first two formant values. However, the experiments in Chapter 5 are equally compatible with the idea that syllables are the basic units, and indeed this hypothesis is supported by data from 15-month-olds. Thiessen and Yee (2010) found that word-level information that put *daw* and *taw* in distinct lexical contexts did not facilitate discrimination of *yad* and *yat*, or of *dee* and *tee*, in the switch task. However, word-level information that illustrated *yad* and *yat* in distinct lexical contexts did facilitate discrimination of these syllables in isolation. These differences may indicate that syllables, rather than phonemes, are the basic units to be categorized, and this would be consistent with the finding that although infants know minimally different words like *ball* and *doll* and can differentiate these in the switch task, they cannot apply this knowledge to novel syllables like *bih* and *dih* (Fennell & Werker, 2003). Alternatively, under a phonemic account, the difficulty in transfer between distinct types of syllables may be related to coarticulatory differences, which are not present in the lexical-distributional model.

These types of assumptions emphasize the fact that this model provides only a starting point to characterize interactive learning of sounds and words. Further research into the mechanisms and representations involved will likely provide evidence against most of the detailed structure of this model. However, the empirical results support the basic idea of interaction between learning processes at the sound and word levels, and it is this interaction that is most central to the model. The remainder of this section outlines potential extensions that might make the model more accurately reflect the problem that infants face in language acquisition.

6.2.2 Phonological alternations and coarticulation

A striking difference between human and model performance was humans' sensitivity to phonological naturalness in Experiment 2, where participants in the LEXICAL condition categorized sounds together when they appeared in complementary distribution, doing so to a greater extent when the alternation was phonologically natural than when it was phonologically unnatural.

In the model of lexical-distributional learning from Chapter 4, phonetic categories were assumed to correspond to the same Gaussian distribution regardless of context, so that the model incorporated no coarticulatory influences or phonological alternations. This simplification prevented the model from showing these types of context-dependent effects. Although it is possible that categorical phonological alternations are learned after phonetic categories have been acquired, through a separate learning algorithm (e.g. Peperkamp, Le Calvez, et al., 2006), this is not likely to be a useful learning mechanism for more gradient types of coarticulatory effects that produce a greater number of acoustic variants. The complexity of the patterns involved (e.g. Hillenbrand et al., 2001) suggests that parametric characterizations of acoustic shifts might be a more effective characterization of gradient coarticulatory effects than separate enumeration of each variant.

Although Dirichlet processes are powerful models of category learning, they do not easily represent sequential dependencies between sounds. Because of the property of exchangeability, any permutation of sounds has the same probability as any other permutation. Hierarchical models can be used to represent bigram statistics (Goldwater et al., 2009) or higher-order structure, such as the organization of sounds into lexical items. However, these simple mechanisms fall short of capturing the complex sequential dependencies found in natural language.

A different class of Bayesian prior, the Markov Random Field (Della Pietra, Della Pietra, & Lafferty, 1997), has been used in recent work to define probabilistic constraints on sequences of speech sounds (Goldwater & Johnson, 2003; Wilson, 2006). Because of the close relation of Markov Random fields to models in formal linguistics (Legendre, Miyata, & Smolensky, 1990), they provide

a promising starting point for incorporating sequencing constraints into models of speech sound categorization. Rather than conditioning an acoustic production only on the category that produced it, the sound can be thought of as the optimal production under a set of interacting constraints.

6.2.3 Phonotactics

In the lexical-distributional model, phonetic categories in lexical items were assumed to be independent of their neighbors, with no phonotactic regularities in the lexicon. This is clearly an incorrect assumption. Consonants and vowels typically alternate within lexical items, and certain clusters of sounds are dispreferred or even excluded from words, with these patterns adhering to language-specific phonotactic restrictions. The model's lack of knowledge of sequencing restrictions in the lexicon likely contributed to its erroneous behavior, causing it to merge similar sounding lexical items because separating them would involve generating the same high probability phonemic sequences several times independently.

Infants are sensitive to phonotactic restrictions in their native language quite early, differentiating between legal and illegal phonotactic sequences (Jusczyk, Friederici, Wessels, & Svenkerud, 1993) and even between legal sequences that have high or low phonotactic probabilities (Jusczyk et al., 1994) by nine months. Recent evidence suggests that this sensitivity to phonotactics may develop even earlier, at six months (Molina & Morgan, to appear).

The assumption of independence was made for purposes of computational simplicity, but it is theoretically possible to incorporate phonotactic constraints into a similar computational framework. Phonotactic constraints might be incorporated by drawing categories according to a hierarchical Dirichlet process, as done for the bigram model in Goldwater et al. (2009), to create sensitivity to pairwise dependencies. However, this would fall short of capturing the rich structure, such as syllable structure and vowel-to-vowel dependencies, that exists across natural languages. Alternatively, a more computationally complex but linguistically more informed option would be to have the generative model for words incorporate interacting phonotactic constraints of the type suggested

in Hayes and Wilson (2008). In each case, the words would retain their linear structure as sequences of phones, but different sequences would have different probabilities of occurrence. This would make the model more likely to hypothesize distinct lexical items for phonotactically probable sequences, whereas it would be more likely to merge phonotactically improbable sequences together with similar sounding words.

6.2.4 Word segmentation

Adults' performance in Experiment 2, which involved word segmentation, contrasted sharply with their performance in Experiment 1, in which words were presented in isolation. It is possible that these differences result from strategies that are specific to word segmentation. While it is not immediately obvious that collapsing *tah* and *taw* together in the LEXICAL condition, but not the NON-LEXICAL condition, would improve the information obtained through transitional probabilities between syllables, a model of simultaneous word segmentation and phonetic category acquisition might help elucidate ways in which the word segmentation task can affect phonetic category learning performance.

Word segmentation could potentially affect phonetic category learning performance by changing the set of words that are available to provide top-down lexical information. For example, infants at 7.5 months would be expected to use monosyllabic words and strong-weak bisyllabic words, but might not be able to use weak-strong bisyllabic words for lexical-distributional learning. In addition, learners might make mistakes in segmenting the input, leading them to hypothesize lexical items that do not occur in the adult lexicon. If the statistics of these subsets of the lexicon are different from the statistics computed over the entire lexicon, this can affect the outcome of the learning process.

One potential way of building this type of joint model is to combine the lexical-distributional model with the unigram word segmentation model from Goldwater et al. (2009). The word segmentation model takes phonemic input and returns a segmentation, and the lexical-distributional

phonetic category learning model takes segmented sequences of acoustic values and returns a set of categories. The two models assume identical structures for the lexicon based on the Dirichlet process. Because each model takes as input the output of the other model, the two models can be used together to jointly recover the categories and segmentation of an unsegmented corpus of acoustic values.

6.2.5 Morphology

Phonetic sequences were assumed in this model to be generated independently for each word, which is not true if a language contains morphological structure. Morphemes, which are reused in different words across the lexicon, could potentially mislead a lexical-distributional learner. For example, in Spanish, verb conjugation patterns like *quiero* 'want-1sg' and *quiere* 'want-3sg' produce sets of minimal pairs that all share the same set of vowels. Similar patterns occur in languages with template morphology, such as Arabic and Hebrew, where the same consonant frame occurs with different sets of vowels in different conjugations of the same word. One potential solution would be to use morphemes, rather than words, as a component of interactive learning. Infants learn about some morphological forms relatively early, recognizing dependencies between the suffix *-ing* and related function words at 18 months (Santelmann & Jusczyk, 1998) and noticing ungrammaticalities related to the suffix *-s* as early as 16 months (Soderstrom, White, Conwell, & Morgan, 2007). Mintz (2004) found that English-learning infants could segment stems from the suffix -ing at 15 months. However, this is later than the time when phonetic category acquisition is first thought to occur. It is possible that infants learning morphologically rich languages show earlier segmentation and recognition of morphemes, parallel to the word segmentation abilities of English-learning infants. Marquis and Shi (2009) found that 11-month-old French-learning infants could recognize inflected nonce verb forms when familiarized with the corresponding nonce verb roots, providing evidence for morphological learning that is a bit earlier, though still slightly later than the time when phonetic categories are first thought to be developing.

A similar type of situation arises in bilingual environments where the two langauges are closely related. Cognates of such languages often differ by a single sound, and this relationship holds consistently across a variety of cognate pairs. This has been suggested as a possible reason why infants who are simultaneously learning Spanish and Catalan show a temporary decline in discrimination around 8 months between the vowel sounds involved in these alternations (Sebastián-Gallés & Bosch, 2009). If correct, this explanation suggests that such developmental patterns of discrimination can be attributed to interactive learning in a learner that has not yet separated the two languages.

6.2.6 Semantics

Whereas this model assumed that learners have no access to semantic information, and thus cannot obtain any disambiguating information from minimal pairs, Yeung and Werker (2009) found evidence that 9-month-old infants are sensitive to pairings of sounds and referents. These infants showed better discrimination of similar sounds when these sounds were consistently paired with different objects than when the sounds were paired with the same set of objects. Although these infants did not necessarily map words to referents in any meaningful way, they showed evidence of modifying their phonetic categorization based on correlations between sounds and objects. One interpretation is that general context, including objects (Yeung & Werker, 2009) and visual information about articulations (Teinonen et al., 2008), can help separate overlapping phonetic categories. Under this view, word-level information might be interpreted as a particularly reliable contextual cue, as it is nearly always present when a sound is heard. Language acquisition is likely to use a combination of these available cues.

Reliance on semantic information, such as information from minimal pairs, is likely to increase over the course of language acquisition as more semantic information becomes available. The function of such semantic cues would be to anchor the category membership of a subset of word tokens. In this way, members of minimal pairs whose referents were known would necessarily be mapped to different lexical items. This type of semi-supervised learning would allow the model to consider only

a subset of the possible partitions of the data, leading the learner to recover the most likely partition of all those that are consistent with the semantic information.

6.3 Empirical extensions

The experimental results suggest that human learning of phonetic categories is more complex than predicted by the lexical-distributional model, with learners behaving differently when words are presented in isolation versus when they are presented in fluent speech. Although human learners use word-level information in each case to constrain their interpretation of phonetic variability, they seem to use this information differently depending on particular aspects of the learning situation. The precise reasons for these differences remain unknown.

One way of addressing this question would be to provide learners with additional cues that favor either a word-level or a phonological interpretation. For example, the consistency of alternations across the lexicon may be important. Given evidence consistent with either a word-level or a vowel harmony interpretation, listeners might be more likely to adopt the vowel harmony interpretation if they hear several words, rather than just two words, that all follow the same pattern. A set of stimuli such as *litah*, *retah*, *mitah*, *gutaw*, *butaw*, and *rotaw* could provide strong evidence that the acoustic differences between *tah* and *taw* are related to the backness of the preceding vowel. This type of evidence might be predicted to induce a phonological interpretation even when words are presented in isolation. Similarly, it will be important to investigate how a word-level interpretation might be obtained when listeners hear words embedded in fluent speech. One immediate follow-up experiment might modify the methods from Experiment 2 to make word segmentation cues more salient. Even a longer familiarization period could help participants recover more of the statistical structure so that they retain information about the critical target words *guta* and *lita*. If participants successfully recover the statistical structure surrounding these target words, it is possible that they may shift their interpretation and view the relevant acoustic diffrerences as phonemic, as though

they were hearing the words in isolation. Alternatively, there is evidence that statistical learning tasks fall short of providing word-level information (Endress & Mehler, 2009). It is possible that other cues to word segmentation, such as stress or duration, would be necessary before participants will transition into a word-level interpretation.

Finally, the experiments presented here demonstrate adults' sensitivity to word-level information, but it is critical to test whether infants are sensitive to these cues at the age when they are first learning phonetic categories. An experiment with 8-month-old infants, roughly parallel to Experiment 1, is currently underway to test infants' sensitivity to word-level information when hearing words in isolation. A positive result in this experiment would provide strong evidence that this potentially useful cue is available to learners while they are still in the process of acquiring the phonetic categories of their native language.

6.4 The importance of interactive learning

The results described in these computational and empirical experiments provide evidence for interactions between learning processes in language acquisition. Interactive learning was shown in principle to lead to a more robust learning outcome than purely distributional learning, given a fixed number of datapoints. Human learners showed influences of higher level information on their phonetic categorization judgments in both experiments, though the specific form of those influences differed between experiments. Moreover, considerations of potential model extensions highlight the idea that it is not just words and sounds that interact during language acquisition. Factors such as semantics, morphology, and phonotactics can provide important information that can potentially influence learning at the word and sound levels.

Previous studies in language acquisition that have investigated how different learning processes interact have typically assumed that learning in different domains occurs sequentially, and have looked at whether learners can use the output of one learning process as input to the next. For

example, Saffran and Wilson (2003) demonstrated that 12-month-old infants could use the words extracted from a segmentation task to learn a rudimentary grammar. Other studies have shown that infants treat the output of statistical learning as potential lexical items, treating them as appropriate insertions in English sentences (Saffran, 2001) and potential object labels (Graf Estes et al., 2007). These studies assumed a unidirectional interaction, but did not examine any potential bidirectional interactions between domains. Several computational models have examined similar types of feed-forward relationships. For example, Christiansen, Onnis, and Hockema (2009) examined how the output of word segmentation could inform syntactic categorization, and Adriaans and Kager (2010) demonstrated that phonotactic learning from a continuous stream of speech could improve word segmentation.

Recently, a number of computational models have begun addressing the issue of bidirectional interactions. Jones, Johnson, and Frank (2010) investigated the interaction between word segmentation and word learning, and Maurits, Perfors, and Navarro (2009) looked at potential benefits of the joint acquisition of word meanings and syntactic word order. To date such investigations have shown a potential benefit of interactive learning in an ideal learner, mirroring the modeling results presented here.

While these computational explorations provide evidence that such interactions help an ideal learner, empirical studies are needed to verify that real learners have the memory and processing resources to make use of such cues. For example, Yurovsky, Yu, and Smith (2010) created an artificial language in which sentences were paired with sets of potential referents in a cross-situational word learning task. Each sentence contained one content word, and that content word was cued either by word-final position, by a frequent preceding word, both, or neither. Participants were tested on segmentation and word-object mapping. Results showed evidence that participants were able to map words to objects in the presence of either cue, but they were not able to do so in the condition where neither cue was present. This suggests that the presence of particular cues that allow for reduced cognitive load may mediate the success of interactive learning in human learners.

Thiessen (2010), using a similar task, provided more optimistic evidence for interactive learning of word segmentation and word meanings. He showed that consistent mappings between auditory and visual stimuli could improve segmentation performance in adults and 20-month-olds performing a simultaneous segmentation and mapping task. This result supports the conclusions of Jones et al. (2010). However, in Thiessen's (2010) experiment, eight month old infants did not show a learning advantage when presented with consistent audiovisual mappings. This suggests a developmental trajectory in which simultaneous acquisition of two layers of structure facilitates learning only once infants have the prerequisite skills to learn both components. Because mappings between words and objects are not learned reliably in the laboratory at eight months, the opportunity to learn these associations simultaneously with a word segmentation task may not benefit learners at this age.

It is important to investigate interactions like these because they can qualitatively change the nature of the learning problem. Whereas separating overlapping categories is difficult for purely distributional learners, it is much easier for interactive learners. Instead, interactive learners might have trouble distinguishing similar sounding words or grouping together the sounds involved in phonological alternations. Research on isolated phenomena, such as phonetic category learning, has produced considerable advances in language acquisition research, but it is important to supplement this type of research by considering the larger picture of language acquisition.

References

Adriaans, F., & Kager, R. (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language*, *62*, 311-331.

Alishahi, A., & Stevenson, S. (2008). A computational model of early argument structure acquisition. *Cognitive Science*, *32*, 789-834.

Alishahi, A., & Stevenson, S. (2010). A computational model of learning semantic roles from child-directed language. *Language and Cognitive Processes*, *25*(1), 50-93.

Anderson, J. L., Morgan, J. L., & White, K. S. (2003). A statistical basis for speech sound discrimination. *Language and Speech*, *46*(2-3), 155-182.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321-324.

Benedict, H. (1979). Early lexical development: comprehension and production. *Journal of Child Language*, *6*, 183-200.

Best, C. T., & McRoberts, G. W. (2003). Infant perception of non-native consonant contrasts that adults assimilate in different ways. *Language and Speech*, *46*(2-3), 183-216.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(3), 345-360.

Boer, B. de, & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, *4*(4), 129-134.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341-345.

Bonatti, L. L., Pea, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations. *Psychological Science*, *16*(6), 451-459.

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science*, *16*(4), 298-304.

Bosch, L., & Sebastián-Gallés, N. (2003). Simultaneous bilingualism and the perception of a language-specific vowel contrast in the first year of life. *Language and Speech*, *46*(2-3), 217-243.

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery.

Machine Learning, *34*, 71-105.

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

Chater, N., & Vitányi, P. (2007). 'Ideal learning' of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology*, *51*, 135-163.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Christiansen, M. H., Onnis, L., & Hockema, S. A. (2009). The secret is in the sound: From unsegmented speech to lexical categories. *Developmental Science*, *12*(3), 388-395.

Curtin, S. (2009). Twelve-month-olds learn novel word-object pairings differing only in stress pattern. *Journal of Child Language*, *36*, 1157-1165.

Della Pietra, S., Della Pietra, V., & Lafferty, J. (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(4), 380-393.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, *39*, 1-38.

Dillon, B., Dunbar, E., & Idsardi, W. (submitted). A single stage approach to learning phonological categories: insights from inuktitut.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*(3968), 303-306.

Emberson, L. L., Liu, R., & Zevin, J. D. (2009). Statistics all the way down: How is statistical learning accomplished using varying productions of novel, complex sound categories? In N. A. Taatgen & H. v. Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Endress, A. D., & Mehler, J. (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, *60*, 351-367.

Fennell, C. T., & Werker, J. F. (2003). Early word learners' ability to access phonetic detail in well-known words. *Language and Speech*, *46*(2), 245-264.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*(2), 209-230.

Ferrier, E. E., & Davis, M. (1973). A lexical approach to the remediation of final sound omissions. *Journal of Speech and Hearing Disorders*, *38*, 126-130.

Fiser, J., & Aslin, R. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 458-467.

Foraker, S., Regier, T., Khetarpal, N., Perfors, A., & Tenenbaum, J. (2009). Indirect evidence and the poverty of the stimulus: The case of anaphoric one. *Cognitive Science*, *33*, 287-300.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578-585.

Fry, D. B., Abramson, A. S., Eimas, P. D., & Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. *Language and Speech*, *5*, 171-189.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE-PAMI*, *6*, 721-741.

Gierut, J. A. (1989). Maximal opposition approach to phonological treatment. *Journal of Speech and Hearing Disorders*, *54*, 9-19.

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*, 103-138.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, *27*(2), 153-198.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2006). Interpolating between types and tokens by estimating power-law generators. *Advances in Neural Information Processing Systems 18*.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21-54.

Goldwater, S., & Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. *Proceedings of the Workshop on Variation within Optimality Theory.*

Goldwater, S., & Johnson, M. (2004). Priors in Bayesian learning of phonological rules. *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON).*

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? *Psychological Science*, *18*(3), 254-260.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Gulian, M., Escudero, P., & Boersma, P. (2007). Supervision hampers distributional learning of vowel contrasts. *ICPhS XVI*.

Hallé, P. A., & Boysson-Bardies, B. d. (1994). Emergence of an early receptive lexicon: Infants' recognition of words. *Infant Behavior and Development*, *17*, 119-129.

Hallé, P. A., & Boysson-Bardies, B. d. (1996). The format of representation of recognized words in infants' early receptive lexicon. *Infant Behavior and Development*, *19*, 463-481.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*, 379-440.

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

Hillenbrand, J. L., Clark, M. J., & Nearey, T. M. (2001). Effects of consonant environment on vowel formant patterns. *Journal of the Acoustical Society of America*, *109*(2), 748-763.

Johnson, M., Griffiths, T. L., & Goldwater, S. (2007). Adaptor grammars: a framework for specifying compositional nonparametric Bayesian models. *Advances in Neural Information Processing Systems 19*.

Jones, B. K., Johnson, M., & Frank, M. C. (2010). Learning words and their meanings from unsegmented child-directed speech. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, 501-509.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*, 1-23.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M. I., & Svenkerud, V. Y. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, *32*, 402-420.

Jusczyk, P. W., Hohne, E. A., & Bauman, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics*, *61*, 1465-1476.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, *39*, 159-207.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the

native language. *Journal of Memory and Language*, *33*, 630-645.

Keating, P. A. (1984). Phonetic and phonological representation of stop consonant voicing. *Language*, *60*(2), 286-319.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307-321.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, *67*(3), 971-995.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, *9*(2), F13-F21.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*(5044), 606-608.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*, 299-321.

Legendre, G., Miyata, Y., & Smolensky, P. (1990). Harmonic grammar: A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. *Technical Report 90-5, Institute of Cognitive Science, University of Colorado.*

Li, P., & Shirai, Y. (2000). *The acquisition of lexical and grammatical aspect.* New York: Mouton de Gruyter.

Lidz, J., Waxman, S., & Freedman, J. (2003). What infants know about syntax but couldn't have learned: experimental evidence for syntactic structure at 18 months. *Cognition*, *89*, B65-B73.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*, 384-422.

MacWhinney, B. (2000). *The CHILDES project.* Mahwah, NJ: Erlbaum.

Mandel, D. R., Jusczyk, P. W., & Pisoni, D. B. (1995). Infants' recognition of the sound patterns of their own names. *Psychological Science*, *6*(5), 314-317.

Mani, N., & Plunkett, K. (2007). Phonological specificity of vowels and consonants in early lexical representations. *Journal of Memory and Language*, *57*, 252-272.

Mani, N., & Plunkett, K. (2010). Twelve-month-olds know their cups from their keps and tups. *Infancy*, *15*(5), 445-470.

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*, 77-80.

Marquis, A., & Shi, R. (2009). The recognition of verb roots and bound morphemes when vowel alternations are at play. In J. Chandlee, M. Franchini, S. Lord, & M. Rheiner (Eds.), *A supplement to the Proceedings of the 33rd Boston University Conference on Language Development.*

Mattys, S. L., & Jusczyk, P. W. (2001a). Do infants segment words or recurring contiguous patterns? *Journal of Experimental Psychology: Human Perception and Performance*, *27*(3), 644-655.

Mattys, S. L., & Jusczyk, P. W. (2001b). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, *78*, 91-121.

Maurits, L., Perfors, A. F., & Navarro, D. J. (2009). Joint acquisition of word order and word reference. In N. A. Taatgen & H. v. Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (p. 1728-1733). Austin, TX: Cognitive Science Society.

Maye, J., & Gerken, L. (2000). Learning phonemes without minimal pairs. In S. C. Howell, S. A. Fish, & T. Keith-Lucas (Eds.), *Proceedings of the 24th Annual Boston University Conference on Language Development* (p. 522-533). Somerville, MA: Cascadilla Press.

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: facilitation and feature generalization. *Developmental Science*, *11*(1), 122-134.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101-B111.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, *12*(3), 369-378.

Meilă, M. (2007). Comparing clusterings – an information based distance. *Journal of Multivariate Analysis*, *98*, 873-895.

Mintz, T. H. (2004). Morphological segmentation in 15-month-old infants. In *Proceedings of the 28th Boston University Conference on Language Development* (p. 363-374). Somerville, MA: Cascadilla Press.

Mintz, T. H., Walker, R. L., Welday, A., & Kidd, C. (in preparation). Infants' universal sensitivity to vowel

harmony and its role in speech segmentation.

Molina, G. C., & Morgan, J. L. (to appear). Sensitivities to native-language phonotactics at 6 months of age. *Poster to be presented at the 2011 Society for Research on Child Development Annual Meeting.*

Narayan, C. R., Werker, J. F., & Beddor, P. S. (2010). The interaction between acoustic salience and language experience in developmental speech perception: evidence from nasal place discrimination. *Developmental Science*, *13*(3), 407-420.

Nazzi, T., Dilley, L. C., Jusczyk, A. M., Shattuck-Hufnagel, S., & Jusczyk, P. W. (2005). English-learning infants' segmentation of verbs from fluent speech. *Language and Speech*, *48*(3), 279-298.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*, 249-265.

Parisien, C., & Stevenson, S. (2010). Learning verb alternations in a usage-based Bayesian model. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Pater, J., Stager, C., & Werker, J. (2004). The perceptual acquisition of phonological contrasts. *Language*, *80*(3), 384-402.

Pearl, L., & Lidz, J. (2009). When domain-general learning fails and when it succeeds: Identifying the contribution of domain specificity. *Language Learning and Development*, *5*, 235-265.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009a). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, *113*, 244-247.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009b). Statistical learning in a natural language by 8-month-old infants. *Child Development*, *80*(3), 674-685.

Peperkamp, S., & Dupoux, E. (2007). Learning the mapping from surface to underlying representations in an artificial language. In J. Cole & J. Hualde (Eds.), *Laboratory phonology 9* (p. 315-338). Berlin: Mouton de Gruyter.

Peperkamp, S., Le Calvez, R., Nadal, J.-P., & Dupoux, E. (2006). The acquisition of allophonic rules: statistical learning with linguistic constraints. *Cognition*, *101*(3), B31-B41.

Peperkamp, S., Pettinato, M., & Dupoux, E. (2003). Allophonic variation and the acquisition of phoneme categories. In B. Beachley, A. Brown, & F. Conlin (Eds.), *Proceedings of the 27th Annual Boston*

University Conference on Language Development (p. 650-661). Somerville, MA: Cascadilla Press.

Peperkamp, S., Skoruppa, K., & Dupoux, E. (2006). The role of phonetic naturalness in phonological rule acquisition. In D. Bamman, T. Magnitskaia, & C. Zaller (Eds.), *Proceedings of the 30th Boston University Conference on Language Development* (p. 464-475). Somerville, MA: Cascadilla Press.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2006). Poverty of the stimulus? a rational approach. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society.*

Perfors, A., Tenenbaum, J. B., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, *37*, 607-642.

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory and Cognition*, *36*(7), 1299-1305.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*(2), 175-184.

Piantadosi, S. T., Goodman, N. D., Ellis, B. A., & Tenenbaum, J. B. (2008). A Bayesian model of the acquisition of compositional semantics. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 421-435.

Port, R. F., & O'Dell, M. L. (1985). Neutralization of syllable-final voicing in German. *Journal of Phonetics*, *13*, 455-471.

Rasmussen, C. E. (2000). The infinite Gaussian mixture model. *Advances in Neural Information Processing Systems 12*, 554-560.

Regier, T., & Gahl, S. (2004). Learning the unlearnable: the role of missing evidence. *Cognition*, *93*, 147-155.

Saffran, J. R. (2001). Words in a sea of sounds: the output of infant statistical learning. *Cognition*, *81*, 149-169.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*, 27-52.

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

Saffran, J. R., & Thiessen, E. D. (2003). Pattern induction by infant language learners. *Developmental Psychology*, *39*(3), 484-494.

Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: Multilevel statistical learning by 12-month-old infants. *Infancy*, *4*(2), 273-284.

Santelmann, L. M., & Jusczyk, P. W. (1998). Sensitivity to discontinuous dependencies in language learners: evidence for limitations in processing space. *Cognition*, *69*, 105-134.

Sato, Y., Sogabe, Y., & Mazuka, R. (2010). Discrimination of phonemic vowel length by Japanese infants. *Developmental Psychology*, *46*(1), 106-119.

Schafer, G., & Plunkett, K. (1998). Rapid word learning by fifteen-month-olds under tightly controlled conditions. *Child Development*, *69*(2), 309-320.

Sebastián-Gallés, N., & Bosch, L. (2009). Developmental shift in the discrimination of vowel contrasts in bilingual infants: is the distributional account all there is to it? *Developmental Science*, *12*(6), 874-887.

Seidl, A., & Buckley, E. (2005). On the learning of arbitrary phonological rules. *Language Learning and Development*, *1*(3-4), 289-316.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13-19.

Soderstrom, M., White, K. S., Conwell, E., & Morgan, J. L. (2007). Receptive grammatical knowledge of familiar content words and inflection in 16-month-olds. *Infancy*, *12*(1), 1-29.

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

Swingley, D. (2005). 11-month-olds' knowledge of how familiar words sound. *Developmental Science*, *8*(5), 432-443.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*, 1566-1581.

Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in

6-month-old infants. *Cognition*, *108*, 850-855.

Teinonen, T., Fellman, V., Näätänen, R., Alku, P., & Huotilainen, M. (2009). Statistical language learning in neonates revealed by event-related brain potentials. *BMC Neuroscience*, *10*, 21.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629-640.

Thiessen, E. D. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language*, *56*(1), 16-34.

Thiessen, E. D. (2010). Effects of simultaneously presented visual information on adults' and infants' auditory statistical learning. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- and 9-month-old infants. *Developmental Psychology*, *39*(4), 706-716.

Thiessen, E. D., & Saffran, J. R. (2007). Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. *Language Learning and Development*, *3*(1), 75-102.

Thiessen, E. D., & Yee, M. N. (2010). Dogs, bogs, labs, and lads: What phonemic generalizations indicate about the nature of children's early word-form representations. *Child Development*, *81*(4), 1287-1303.

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, *10*(2), 172-175.

Toro, J. M., Nespor, M., Mehler, J., & Bonatti, L. L. (2008). Finding words and rules in a speech stream. *Psychological Science*, *19*(2), 137-144.

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

Trehub, S. E. (1976). The discrimination of foreign speech contrasts by infants and adults. *Child Development*, *47*(2), 466-472.

Trubetzkoy, N. S. (1939). *Grundzüge der Phonologie.* Göttingen: Vandenhoeck und Ruprecht.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, *104*, 13273-13278.

Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M., & Stager, C. L. (1998). Acquisition of word-object associations by 14-month-old infants. *Developmental Psychology*, *34*(6), 1289-1309.

Werker, J. F., Fennell, C. T., Corcoran, K. M., & Stager, C. L. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy*, *3*(1), 1-30.

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*, 49-63.

White, K. S., Peperkamp, S., Kirk, C., & Morgan, J. L. (2008). Rapid acquisition of phonological alternations by infants. *Cognition*, *107*(1), 238-265.

Wilson, C. (2003). Experimental investigation of phonological naturalness. *Proceedings of WCCFL 22*.

Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*, *30*, 945-982.

Wonnacott, E., Newport, E. L., & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology*, *56*, 165-209.

Woodward, A. L., Markman, E. M., & Fitzsimmons, C. M. (1994). Rapid word learning in 13- and 18-month-olds. *Developmental Psychology*, *30*(4), 553-566.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245-272.

Yeung, H. H., & Werker, J. F. (2009). Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition*, *113*(2), 234-243.

Yoshida, K. A., Fennell, C. T., Swingley, D., & Werker, J. F. (2009). Fourteen month-old infants learn similar sounding words. *Developmental Science*, *12*(3), 412-418.

Yoshida, K. A., Pons, F., Maye, J., & Werker, J. F. (2010). Distributional phonetic learning at 10 months of age. *Infancy*, *15*(4), 420-433.

Yurovsky, D., Yu, C., & Smith, L. B. (2010). Statistical speech segmentation and word learning in parallel. In K. Franich, K. M. Iserman, & L. L. Keil (Eds.), *Proceedings of the 34th Boston University Conference on Language Development* (p. 491-502). Somerville, MA: Cascadilla Press.

Appendix A

Likelihood computation

Here a derivation is given for the likelihood terms used in the Gibbs sampling algorithms for the lexical-distributional model and the IMM. Computation of the likelihood term $p(w_{kj}|z, w_{-kj}, l)$ involves marginalizing over the mean μ and covariance Σ of the phonetic category l_{kj},

$$p(w_{kj}|z, w_{-kj}, l) = \int p(\Sigma|z, w_{-kj}, l) \int p(\mu|\Sigma, z, w_{-kj}, l) p(w_{kj}|\mu, \Sigma) d\mu d\Sigma \tag{A.1}$$

where a dependence on hyperparameters μ_0, Σ_0, and ν_0 has been suppressed for readability. Because phonetic detail is generated independently each time a given phoneme is uttered, the last term can be factored into the probabilities of the individual speech sounds w_{ij} in the set w_{kj}, so the expression on the right hand side becomes

$$\int p(\Sigma|z, w_{-kj}, l) \int p(\mu|\Sigma, z, w_{-kj}, l) \prod_{i=1}^{n} p(w_{ij}|\mu, \Sigma) d\mu d\Sigma \tag{A.2}$$

Under the assumptions of the generative model, if μ_c, ν_c, and Σ_c are current estimates of category parameters as defined in Equations 4.4-4.6, the first term in the above equation has an inverse Wishart distribution $IW(\nu_c, \Sigma_c)$, the second is $N\left(\mu_c, \frac{\Sigma}{\nu_c}\right)$, and each of the terms inside the product is $N(\mu, \Sigma)$. Expanding these terms yields

$$\begin{aligned}
& \int \frac{|\frac{\Sigma_c}{2}|^{\frac{\nu_c}{2}}}{\Gamma_d(\frac{\nu_c}{2})} \frac{e^{-\frac{1}{2}tr(\Sigma_c\Sigma^{-1})}}{|\Sigma|^{\frac{\nu_c+d+1}{2}}} \int \frac{1}{(2\pi)^{\frac{d}{2}}|\frac{\Sigma}{\nu_c}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mu-\mu_c)^T\left(\frac{\Sigma}{\nu_c}\right)^{-1}(\mu-\mu_c)} \\
& \quad \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(w_{ij}-\mu)^T\Sigma^{-1}(w_{ij}-\mu)} d\mu d\Sigma \\
= & \int \frac{|\frac{\Sigma_c}{2}|^{\frac{\nu_c}{2}}}{\Gamma_d(\frac{\nu_c}{2})} \frac{e^{-\frac{1}{2}tr(\Sigma_c\Sigma^{-1})}}{|\Sigma|^{\frac{\nu_c+d+1}{2}}} \int \frac{1}{(2\pi)^{\frac{d}{2}}|\frac{\Sigma}{\nu_c}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mu-\mu_c)^T\left(\frac{\Sigma}{\nu_c}\right)^{-1}(\mu-\mu_c)} \\
& \quad \frac{1}{(2\pi)^{\frac{dn}{2}}|\Sigma|^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n}(w_{ij}-\bar{w}_{kj})^T\Sigma^{-1}(w_{ij}-\bar{w}_{kj}) - \frac{1}{2}(\bar{w}_{kj}-\mu)^T\left(\frac{\Sigma}{n}\right)^{-1}(\bar{w}_{kj}-\mu)} d\mu d\Sigma \\
= & \int \frac{|\frac{\Sigma_c}{2}|^{\frac{\nu_c}{2}}}{\Gamma_d(\frac{\nu_c}{2})} \frac{e^{-\frac{1}{2}tr(\Sigma_c\Sigma^{-1})}}{|\Sigma|^{\frac{\nu_c+d+1}{2}}} \frac{e^{-\frac{1}{2}\sum_{i=1}^{n}(w_{ij}-\bar{w}_{kj})^T\Sigma^{-1}(w_{ij}-\bar{w}_{kj})}}{(2\pi)^{\frac{d(n-1)}{2}}|\Sigma|^{\frac{n-1}{2}}n^{\frac{d}{2}}} \\
& \quad \int \frac{1}{(2\pi)^{\frac{d}{2}}|\frac{\Sigma}{\nu_c}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mu-\mu_c)^T\left(\frac{\Sigma}{\nu_c}\right)^{-1}(\mu-\mu_c)} \\
& \quad \frac{1}{(2\pi)^{\frac{d}{2}}|\frac{\Sigma}{n}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\bar{w}_{kj}-\mu)^T\left(\frac{\Sigma}{n}\right)^{-1}(\bar{w}_{kj}-\mu)} d\mu d\Sigma
\end{aligned} \tag{A.3}$$

The solution to the inner integral is a normal distribution centered at μ_c whose variance is given by

the sum of the variances, $\frac{\Sigma}{\nu_c} + \frac{\Sigma}{n}$. The expression therefore becomes

$$\begin{aligned}
&\int \frac{|\frac{\Sigma_c}{2}|^{\frac{\nu_c}{2}}}{\Gamma_d(\frac{\nu_c}{2})} \frac{e^{-\frac{1}{2}tr(\Sigma_c\Sigma^{-1})}}{|\Sigma|^{\frac{\nu_c+d+1}{2}}} \frac{e^{-\frac{1}{2}\sum_{i=1}^{n}(w_{ij}-\bar{w}_{kj})^T\Sigma^{-1}(w_{ij}-\bar{w}_{kj})}}{(2\pi)^{\frac{d(n-1)}{2}}|\Sigma|^{\frac{n-1}{2}}n^{\frac{d}{2}}} \\
&\qquad \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma(\frac{n+\nu_c}{n\nu_c})|^{\frac{1}{2}}} e^{-\frac{1}{2}(\bar{w}_{kj}-\mu_c)^T\left(\Sigma(\frac{n+\nu_c}{n\nu_c})\right)^{-1}(\bar{w}_{kj}-\mu_c)} d\Sigma \\
= &\int \frac{|\frac{\Sigma_c}{2}|^{\frac{\nu_c}{2}}}{\Gamma_d(\frac{\nu_c}{2})(2\pi)^{\frac{dn}{2}}\frac{n+\nu_c}{\nu_c}^{\frac{d}{2}}|\Sigma|^{\frac{\nu_c+d+n+1}{2}}} \\
&\qquad e^{-\frac{1}{2}\left[tr(\Sigma_c\Sigma^{-1})+tr\left(\sum_{i=1}^{n}(w_{ij}-\bar{w}_{kj})^T\Sigma^{-1}(w_{ij}-\bar{w}_{kj})\right)+tr((\bar{w}_{kj}-\mu_c)^T\left(\Sigma(\frac{n+\nu_c}{n\nu_c})\right)^{-1}(\bar{w}_{kj}-\mu_c)\right)\right]} d\Sigma \\
= &\frac{|\frac{\Sigma_c}{2}|^{\frac{\nu_c}{2}}}{\Gamma_d(\frac{\nu_c}{2})(2\pi)^{\frac{dn}{2}}\frac{n+\nu_c}{\nu_c}^{\frac{d}{2}}} \int \frac{1}{|\Sigma|^{\frac{\nu_c+d+n+1}{2}}} \\
&\qquad e^{-tr\left(\frac{1}{2}\left[\Sigma_c+\sum_{i=1}^{n}(w_{ij}-\bar{w}_{kj})(w_{ij}-\bar{w}_{kj})^T+\left(\frac{n\nu_c}{n+\nu_c}\right)(\bar{w}_{kj}-\mu_c)(\bar{w}_{kj}-\mu_c)^T\right]\Sigma^{-1}\right)} d\Sigma
\end{aligned} \tag{A.4}$$

This integral has the form $\int |V|^{-k-\frac{d+1}{2}} e^{-tr(AV^{-1})} dV$, an integral which is equal to $|A|^{-k}\Gamma_d(k)$. The entire expression can therefore be written as

$$\begin{aligned}
&\frac{|\frac{\Sigma_c}{2}|^{\frac{\nu_c}{2}}}{\Gamma_d(\frac{\nu_c}{2})(2\pi)^{\frac{dn}{2}}\frac{n+\nu_c}{\nu_c}^{\frac{d}{2}}} \left|\frac{1}{2}[\Sigma_c + \sum_{i=1}^{n}(w_{ij}-\bar{w}_{kj})(w_{ij}-\bar{w}_{kj})^T + \right. \\
&\qquad \left.\left(\frac{n\nu_c}{n+\nu_c}\right)(\bar{w}_{kj}-\mu_c)(\bar{w}_{kj}-\mu_c)^T]\right|^{-\frac{\nu_c+n}{2}} \Gamma_d\left(\frac{\nu_c+n}{2}\right) \\
= &\frac{\Gamma_d(\frac{\nu_c+n}{2})|\Sigma_c|^{\frac{\nu_c}{2}}}{\Gamma_d(\frac{\nu_c}{2})\pi^{\frac{dn}{2}}\frac{n+\nu_c}{\nu_c}^{\frac{d}{2}}} \left|\Sigma_c + \sum_{i=1}^{n}(w_{ij}-\bar{w}_{kj})(w_{ij}-\bar{w}_{kj})^T + \right. \\
&\qquad \left.\left(\frac{n\nu_c}{n+\nu_c}\right)(\bar{w}_{kj}-\mu_c)(\bar{w}_{kj}-\mu_c)^T\right|^{-\frac{\nu_c+n}{2}}
\end{aligned} \tag{A.5}$$

Equation A.5 is the general form for the likelihood. In the special case where $n = 1$, however, it can be simplified further. The form in this case is

$$\frac{\Gamma_d(\frac{\nu_c+1}{2})|\Sigma_c|^{\frac{\nu_c}{2}}}{\Gamma_d(\frac{\nu_c}{2})\pi^{\frac{d}{2}}\frac{\nu_c+1}{\nu_c}^{\frac{d}{2}}} \left|\Sigma_c + \left(\frac{\nu_c}{\nu_c+1}\right)(w_{ij}-\mu_c)(w_{ij}-\mu_c)^T\right|^{-\frac{\nu_c+1}{2}} \tag{A.6}$$

The identity $|X + ab^T| = |X|(1 + b^T X^{-1} a)$ can be applied, yielding

$$\begin{aligned}
&\frac{\Gamma_d(\frac{\nu_c+1}{2})|\Sigma_c|^{\frac{\nu_c}{2}}}{\Gamma_d(\frac{\nu_c}{2})\pi^{\frac{d}{2}}\frac{\nu_c+1}{\nu_c}^{\frac{d}{2}}} |\Sigma_c|^{-\frac{\nu_c+1}{2}} \left(1 + (w_{ij}-\mu_c)^T\left[\Sigma_c\left(\frac{\nu_c+1}{\nu_c}\right)\right]^{-1}(w_{ij}-\mu_c)\right)^{-\frac{\nu_c+1}{2}} \\
= &\frac{\Gamma(\frac{\nu_c+1}{2})}{\Gamma(\frac{\nu_c+1-d}{2})} \left|\pi\Sigma_c\left(\frac{\nu_c+1}{\nu_c}\right)\right|^{-\frac{1}{2}} \left(1 + (w_{ij}-\mu_c)^T\left[\Sigma_c\left(\frac{\nu_c+1}{\nu_c}\right)\right]^{-1}(w_{ij}-\mu_c)\right)^{-\frac{\nu_c+1}{2}}
\end{aligned} \tag{A.7}$$

which is a multivariate t-distribution with mean μ_c, variance $\Sigma_c \left(\frac{\nu_c+1}{\nu_c}\right)$, and degrees of freedom $\nu_c + 1 - d$.