FROM PERCEPTUAL LEARNING TO SPEECH PRODUCTION: GENERALIZING PHONOTACTIC PROBABILITIES IN LANGUAGE ACQUISITION

by

Peter T Richtsmeier

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF LINGUISTICS

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2008

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we have read the dissertation

prepared by Peter T Richtsmeier

entitled From Perceptual Learning to Speech Production: Generalizing Phonotactic Probabilities in Language Acquisition

and recommend that it be accepted as fulfilling the dissertation requirement for the

degree of Philosophy

____ Date: 6/13/2008
Diane Ohala

____ Date: 6/13/2008
LouAnn Gerken

____ Date: 6/13/2008
Andrew Lotto

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

____ Date: 6/13/2008
Dissertation Director: Diane Ohala

____ Date: 6/13/2008

Dissertation Director: LouAnn Gerken

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Peter T Richtsmeier

ACKNOWLEDGEMENTS

I acknowledge to you, the reader, this list of advisors, assistants, colleagues, friends, and family, who made this dissertation possible.

- My committee, Diane Ohala, LouAnn Gerken, and Andrew Lotto, for assistance in preparing the design, executing the experiments, and writing the manuscript
- Diane, for detailed and insightful feedback, and for sticking with me from year one
- LouAnn, for bringing me in on the project that led to this dissertation
- Lisa Goffman, for her friendship, her sound advice and good ideas, and for opening up future possibilities related to this work
- Mike Hammond, for introducing me to LaTeX , and for help in more ways that I can recount
- Juliet Minton for help recruiting and running participants, and for some seriously awesome baked goods
- Jordan Brewer, Lynnika Butler, Claire Fischer, Amy Fountain, Karen Gerberding, Andrea González, Kara Hawthorne, Evelyn Jaramillo, Amy LaCross, Brittany Lindsay, Juliet Minton, Rabiah Muhammad, Erin O'Bryan, Polly O'Rourke, Sumayya Racy, and Michelle Sandoval, for serving as talkers in my experiments
- Brianna McMillan, Lauren Akif, and Claire Fischer, for help with the reliability analyses
- Tiffany Hogan and Michael Vitevitch, for help with stimuli preparation
- Roland Hancock, for help with statistics
- Adam Albright, for inspiration and assistance with a reference list for the type frequency literature
- Natasha Warner, for assistance with a reference list for the exemplar model literature
- Mary Beckman, Jan Edwards, Ben Munson, and Janet Pierrehumbert, for inspiration
- Merrill Garrett, Vic Ferreira, Karen Stromswold, and members of the Sound Minds phonology group, for their insightful comments
- My family, my parents Tom and Jean Richtsmeier, my sisters Mary and Teresa, my brothers Sam and Tony, my nephews Xavier and Harlen, my stepsister Toni, and all of my extended family, for love and support
- Lastly, my wife, Erin O'Bryan, for teaching me a set of valuable research skills, for moral support throughout difficult periods, and for being my true love

DEDICATION

I dedicate this work to my father, Thomas Elliott Richtsmeier, and to my wife, Erin Leigh O'Bryan. You both exemplify scholarship with your passion for knowledge and truth. I hope my career as a scientist will inspire others as you have inspired me.

Printed in $\LaTeX$

TABLE OF CONTENTS

TABLE OF CONTENTS – *Continued*

TABLE OF CONTENTS – *Continued*

LIST OF FIGURES

LIST OF FIGURES – *Continued*

LIST OF TABLES

ABSTRACT

Phonotactics are the restrictions on sound sequences within a word or syllable. They are an important cue for speech segmentation and a guiding force in the creation of new words. By studying phonotactics, we stand to gain a better understanding of why languages and speakers have phonologies. Through a series of four experiments, I will present data that sharpen our theoretical and empirical perspectives of what phonotactics are and how they are acquired.

The methodology is similar to that used in studies of infant perception: children are familiarized with a set of words that contain either a few or many examples of a phonotactic sequence. The participants here are four-year-olds, and the test involves producing a target phonotactic sequence in a new word. Because the test words have not been encountered before, children must generalize what they learned in the familiarization phase and apply it to their own speech. By manipulating the phonetic and phonological characteristics of the familiarization items, we can determine which factors are relevant to phonotactic learning. In these experiments, the phonetic manipulation was the number of talkers who children heard produce a familiarization word. The phonological manipulation was the number of familiarization words that shared a phonotactic pattern.

The findings include instances where learning occurs and instances where it does not. First, the data show that the well-studied correlation between phonotactic probability and production accuracy in child speech can be attributed, at least partly to perceptual learning, rather than a practice effect attributable to repeated articulation. Second, the data show that perceptual learning is a process of abstraction and learning about those abstractions. It is not about making connections between stored, unelaborated exemplars because learning from the phonetic manipulation alone was insufficient for a phonotactic pattern to generalize. Furthermore, perceptual learning is not about reorganizing pre-existing symbolic knowledge, because learning from words alone is insufficient. I argue that a model

which learns abstract word-forms from direct phonetic experience, then learns phonotactics from the abstract word-forms, is the most parsimonious explanation of phonotactic learning.

CHAPTER 1

INTRODUCTION

1.1 The Broader Context

This dissertation is about learning phonotactic probabilities. The term *phonotactic probability*, also referred to as phonotactic frequency, is the likelihood that a sound or sound sequence occurs in a particular part of a word or syllable (Kenstowicz, 1994). The study of phonotactic probabilities has grown out of the study of phonotactics, which is the study of allowable sound sequences in a language. The term “allowable” has generally been used to refer to categorical knowledge, such as the fact that adult English speakers accept the onset /bl/ as possible in English, as in the nonsense word *blick*, but do not accept the onset /bn/ in *bnick* . In other words, the traditional definition of phonotactics is the study of which sounds sequences are in a language versus which sound sequences are out. Phonotactic probabilities, in contrast, are more or less frequent relative to one another, and represent a gradient spanning from the least frequent to the most frequent sound sequences in a language. For example, the probability of /bl/ as an onset is relatively high because it occurs in many English words (*blue*, *black*, *blend*, *blouse*, *bleary*, etc.). The onset /sf/ occurs in relatively few words, however (*sphere*, *sphinx*, *sphincter*, etc.). This frequency gradient corresponds to the relative acceptability of the two onsets: adults generally prefer nonsense words like *blick* to words like *sphick* (Hammond, 2004).

Phonotactics are important because they tell us about how linguistic knowledge is represented—that is, they tell us about grammar—and they tell us about how individuals learn to perceive and produce speech. With respect to knowledge and grammar, the study of phonotactics has shown us that grammatical knowledge includes the context that sounds appear in. Contextual knowledge of sounds can be categorical, as in the distinction between /bl/ and /bn/ onsets, or it can be probabilistic, as in the distinction between /bl/ and /sf/. With respect to speech perception and production, work with infants has shown that sound sequences are important cues to word boundaries (Friederici & Wessels, 1993; Saffran, Aslin, & Newport, 1996). That is, infants must learn to find words in a fluid speech stream (Mehler, Dupoux, & Segui, 1990; Woodward & Aslin, 1990), and infants use phonotactic probabilities to accomplish this task (Mattys & Jusczyk, 2000; Mattys, Jusczyk, Luce, & Morgan, 1999). Phonotactics are also relevant to word learning. A series of studies by Holly Storkel and colleagues (Storkel, 2001, 2004; Storkel & Morrisette, 2002; Storkel & Rogers, 2000) has shown that high probability phonotactics support word learning. Children learn new words more readily when they are composed of common sound sequences (for additional discussion and examples of phonotactic effects in infant and child development, see the literature review in the next chapter).

Researchers studying phonological acquisition have found that phonotactic effects are common in child speech. For example, children seem more adept at producing a high probability sequence, such as the consonant sequence /ft/ in the nonsense word /moftın/ compared to a low probability sequence, such as the /fk/ in /mofkən/ (Beckman & Edwards, 1999; Edwards, Beckman, & Munson, 2004; Munson, 2001; Zamuner, Gerken, & Hammond, 2004). These studies have shown that children produce more probable sequences with a higher degree of accuracy and fluency than less probable sequences. We can conclude from the child and infant research that learning phonotactics is a critical component to learning a language, so it is natural that we work to better our understanding of them (Jusczyk, 1997).

Returning to the issue of grammar, phonotactics have played an important role in shifting linguistic research from a focus on all-or-nothing categorical knowledge like that discussed in Chomsky and Halle (1968) and Kenstowicz (1994) to the study of finer-grained and nuanced knowledge. Hammond (2004), for example, discusses the fact that English speakers' preference for /bl/ onsets relative to /sf/ onsets is surprising from the perspective of categorical phonotactics, given that both onsets occur in English words (in *blinks* and *sphinx*, for example). Similarly, Munson (2001) showed that adult English speakers have a preference for the /mp/ cluster of /fæmpət/ compared to the /ft/ cluster of /fæftət/. These and similar findings throughout the literature are generally accepted to reflect the fact that much of linguistic knowledge is gradient, rather than categorical.

Why are phonotactics probabilistic? One answer is that phonotactics are abstract or symbolic representations of statistical patterns present across the set of word-forms stored by a speaker, or the speaker's lexicon (Bailey & Hahn, 2001; Coleman & Pierrehumbert, 1997; Greenberg & Jenkins, 1964; J. Ohala & Ohala, 1986). This claim has been motivated by research on people's judgments about the acceptability or well-formedness of phonotactic sequences (Coleman & Pierrehumbert, 1997; Albright, 2007), and the consistent finding that people judge high probability sequences to be more acceptable than low probability sequences. Because well-formedness judgments are thought to reflect the grammar, the implication is that the grammar arises from general processes such as word learning[1].

Positing the lexicon as the source of phonotactic probabilities still leaves open several possible descriptions of the grammar, however. Explanations differ widely and range from proposals that advocate a rule-based view of phonotactic probabilities (Albright, 2007), a constraint-based view (Hammond, 2003, 2004; Kisseberth, 1970), and a prototype-based view (Pierrehumbert, 2003b). All of these approaches draw on theoretical mechanisms developed in the generative grammar framework (Chomsky & Halle, 1968; Prince & Smolensky, 1993), but formalize these mechanisms in quite different ways. The approaches offered by Albright (2007) and Pierrehumbert (2003b), as well as the representations that they adopt, are discussed in greater detail in Section 2.4.1 in Chapter 2.

Another answer to the question of why phonotactics are probabilistic is that phonotactics reflect the accumulation of direct experience, rather than generalizations made about

[1] Phonotactic gradients have also been found in people's judgments of sequences that do not appear in their native langauge. For example, Albright (2007) shows that English speakers prefer the /pw/ onset of /pwʌdz/ to the /bz/ onset of /bzarʃk, although neither onset is typical of English words. With respect to acquisition, Pertz and Bever (1975) show that adolescents have a greater sensitivity to cross-linguistic phonotactic constraints than younger children. These authors argue that universal knowledge of phonotactics develops through the interaction of linguistic experience and a hardwired acquisition process. Finally, Berent and colleagues have shown that speakers of a number of languages treat unknown phonotactic sequences consistently. For example, Berent, Lennertz, Jun, Moreno, and Smolensky (2008) showed that Korean speakers consistently parse novel consonant clusters according to predictions made by the Sonority Hierarchy (Jespersen, 1904), although no consonant clusters exist in Korean. In other words, there is no positive evidence for the Sonority Hierarchy in that language. Together, these studies provide convincing evidence that certain aspects of phonotactic patterning are not learned. These findings are not incompatible with the learning-based tack taken here, but their treatment is beyond the scope of this dissertation, and so they will not be covered. See Albright, 2007, for further discussion of the issue.

the lexicon. A number of psycholinguistic studies have shown that words are associated in memory, and that these associations are highly correlated with phonotactic probability (Bailey & Hahn, 2001; Charles-Luce & Luce, 1995; Frauenfelder & Schreuder, 1992; Gathercole & Martin, 1996; Landauer & Streeter, 1973; Luce, Pisoni, & Goldinger, 1990; Vitevitch & Luce, 1998, 1999). This fact suggests that phonotactic probabilities may be a reflection of the storage of words, rather than the existence of abstract or symbolic knowledge. In other words, phonotactic knowledge may not exist in and of itsef, but may instead be the by-product of more basic processes of perception and memory. This is precisely the mechanism described by Bailey and Hahn (2001), who combined an exemplar-based approach to storage (Hintzman, 1986; Nosofsky, 1986) with models of lexical neighborhood activation (Luce, 1986; McClelland & Elman, 1986; Norris, 1994; Vitevitch & Luce, 1998) to provide a coherent framework for explaining phonotactic probabilities without explicit phonotactic knowledge[2]. The tension between explanations that rely on direct experience (Bailey & Hahn, 2001) versus those that refer to a lexicon of abstract wordforms (Albright, 2007; Pierrehumbert, 2003b), often referred to as the types versus tokens debate (cf. Bailey & Hahn, 2001; Bybee, 1995; Albright & Hayes, 2003; Albright, 2007; Pierrehumbert, 2003b for additional discussion), is at the heart of the debate about linguistic representations, which plays heavily into our understanding of what phonotactics are and why they are probabilistic.

What phonotactic probabilities are poses another important empirical question, namely, *how* phonotactic probabilities are learned. Past research has largely speculated about the learning mechanism responsible for phonotactic knowledge: phonotactic probabilities could be learned via perceptual experience, as infant studies suggest (Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993; Jusczyk, Luce, & Charles-Luce, 1994), or through articulatory practice of the sound sequences (Edwards et al., 2004; Messum, 2007), or from some combination of the two. For example, we might hypothesize that

[2] Bailey and Hahn's findings suggest that their exemplar model is not adequate to explain phonotactic probabilities, although they conclude that future exemplar models may achieve explanatory adequacy. Here and in subsequent chapters, I will discuss Bailey and Hahn's model as if it could stand alone. I have chosen to discuss the model in this way because of Bailey and Hahn's conclusion that future research may provide further support for the model, and because it is the only model of its kind that I am aware of.

children in the Beckman and Edwards (1999) study were more accurate at producing /ft/ than /fk/ because they had heard the former sequence more often, appearing in words like *soft* and *lift* in the speech of people around them, but heard few if any words containing /fk/. An alternative explanation is that children have produced *soft*, *lift*, and other /ft/ words enough times (cf. the expressive vocabulary norms given by Dale & Fenson, 1996) to have developed a certain level of articulatory expertise, and that expertise is what separates /ft/ and /fk/. Finally, we can hypothesize that both perceptual and articulatory knowledge of a phonotactic sequence combine to improve a child's accuracy in producing that sequence.

To conclude, the general context of the dissertation can be summed up as follows: phonotactic probability, or the likelihood of a sound sequence in its phonological context in a language. Probabilistic phonotactic effects can be seen in child speech—children are more accurate when producing high probability phonotactic sequences compared to low probability sequences. One explanation that has been offered for why phonotactics are probabilistic is that phonotactic knowledge is predicated on word learning. In other words, the probability of a phonotactic sequence depends directly on the number of words in the language that contain that sequence. By adopting this explanation, we must still address several important empirical questions, however. First, we should ask what psychological mechanism is responsible for the creation of those representations. Second, we should ask what the representations of phonotactic probabilities are that word learning creates. In the next section, I introduce the methodology for studying phonotactic probabilities. This methodology harnesses one possible answer to the first question and provides an answer to the second question.

1.2 The Narrow Methodological Focus

The dissertation is composed of four experiments designed under one coherent methodology. In every experiment, the target population is children about four years of age. Four-year-olds are an ideal population because they are still relatively unsophisticated in their speech production, but are also far enough along in development that the experiment can be explained to them.

With respect to materials, I focus on a phonotactic unit commonly referred to as the biphone. A *biphone* is a sequence of two sounds in some position in a word, whereas a *phone* is just a single sound. The examples given in the introductory paragraph above, /bl/, /sf/, and /bn/, are all examples of biphones. More specifically, they are 2-consonant biphones, or clusters, which are the particular biphones that I also focus on. In language acquisition research, consonant clusters have consistently been used to study phonological (Locke, 1983; McLeod, Doorn, & Reed, 2001; D. K. Ohala, 1999; Olmsted, 1971; Prather, Hedrick, & Kern, 1975; Smit, 1993) and phonotactic (Beckman & Edwards, 1999; Edwards et al., 2004; Munson, 2001; Gerken et al., 2006) development. Clusters have the advantage of being a relatively coherent phonotactic unit while spanning a wide range of phonotactic probabilities.

The methodology is dependent on the perceptual learning explanation for how children acquire phonotactic probabilities: that is, the experiments are designed to test what children can learn from a perceptual input. The procedure involves 'familiarization followed by testing.' The clusters are first introduced to children in perception, then we look for evidence of perceptual learning in children's speech production. Some existing evidence (Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999; Gerken et al., 2006; Wang, Jongman, & Sereno, 2003) points to the perceptual learning mechanism as being responsible for production accuracy. More detailed coverage of the potential mechanisms behind phonotactic learning and the likelihood of a perceptual learning mechanism are given in Chapter 2. As we will see, to the extent that perceptual learning effects have been found in production, we have clear evidence for a perceptual learning mechanism. In this way, the methodology harnesses one of the possible answers to the question of how phonotactic probabilities are learned.

The procedure relies heavily on artificial language learning paradigms (Chambers, Onishi, & Fisher, 2003; Gómez & Gerken, 1999; Marcus, Vijayan, Bandi-Rao, & Vishton, 1999; Reber, 1963, 1967, see Gómez & Gerken, 2000 for a review) that have been used in infant studies. As discussed in the literature review in the next chapter, infants in an artificial language learning experiment are familiarized with a set of linguistic patterns during a familiarization phase in which they listen passively. They are then tested for learning based on a measurement of their looking times for learned versus unlearned sequences (Jusczyk & Aslin, 1995). This methodology allows for the real-time creation of linguistic knowledge. In the experiments presented here, the target population is speaking children and the linguistic pattern of interest is phonotactics. The test, however, involves speaking rather than looking.

The effects of perceptual learning are analyzed in two dependent measures: accuracy and production latency. Accuracy is a transcription based measure which codes whether children were able to produce the consonants of the target cluster accurately, inaccurately, or not at all. Production latencies are similar to reaction times. They are the time from the end of the target word to the beginning of the child's production.

The particular focus of each experiment can be described as follows. Experiment 1 tests whether a phonotactic pattern can be generalized in a production task based on perceptual familiarization of that pattern in word-tokens spoken by multiple talkers. Experiments 2A and 2B test whether a phonotactic pattern can be generalized based on familiarization with multiple word-types spoken by multiple talkers. Finally, Experiment 3 tests whether multiple word-types alone in the familiarization can lead to generalization. This independent manipulation of token variability and type frequency will allow us to answer questions about what representations underly probabilistic phonotactic effects, such as differences in production accuracy.

It should be noted that the critical phonotactic probabilities in this experiment are artificial ones. English-based phonotactics are controlled in the experiment, and children are asked to produce both high and low probability phonotactic sequences, but English phonotactics cannot be created within an experimental setting. To create phonotactic effects, a manipulation referred to as *Experimental Frequency* will be used in all four experiments, which is a manipulation of the number of times that a phonotactic sequence is heard *within the experiment*. The means by which we can infer an effect of Experimental Frequency is by measuring the speed and accuracy of children's productions. These are the same measurements used in studies of English phonotactics (Beckman & Edwards, 1999; Edwards et al., 2004; Munson, 2001; Zamuner et al., 2004), so when an effect of Experimental Frequency is found, it provides the same type of evidence that led to the conclusion that English phonotactics affect child speech. This methodology therefore provides a means for studying how English phonotactic probability affects arise, and so it adds a new level of external validity to the artificial language learning paradigm.

1.3 The Relevance of the Dissertation

The primary goal of this dissertation is to add to the base of scientific knowledge of phonotactics and phonotactic probabilities in particular, and to test some of the key properties of existing models of phonotactic learning. For these purposes, this dissertation presents four studies of perceptual learning conducted with 4-year-olds to probe the nature of the representation of such information in the grammar. The perceptual learning methodology has the advantage of allowing for a comparison of token- and type-based learning, which can be independently manipulated in a perceptual input. Token-based and type-based models make different predictions about how phonotactic effects should come into existence. One such effect, that children are more accurate to produce high probability phonotactic sequences (Beckman & Edwards, 1999; Edwards et al., 2004; Munson, 2001; Zamuner et al., 2004), can be used to compare the relative contributions of phonetic variability and type frequency as they contribute to children's production accuracy.

This dissertation makes unique emprical and theoretical contributions to the study of phonotactic probabilities, specifically, and language, generally. By using a design in which perceptual learning is directly implicated in speech production, the findings tie together two different research areas: infant perception studies and child production studies. Although researchers in both areas are in agreement about the relevance of linguistic experience, they often draw upon different mechanisms to explain their presence, or even conflate potential mechanisms. By testing one mechanism that links both areas, namely, perceptual learning, the experiments show that infants and older children can and do learn about language in a fundamentally similar way.

The dissertation also advances the precision of our theoretical description of phonotactics because the design allows for a distinction between phonotactic learning that requires no abstract knowledge and phonotactic learning that does. This is made possible by the manipulation of multiple potential levels of abstraction. The methodology examines learned or bottom-up generalizations, but manipulates the presence of a level of token variability that has been implicated for knowledge of words, that is, talker variability (Houston & Jusczyk, 2003), and a level of word variability implicated for knowledge of phonotactics, that is, type frequency (Albright, 2007; Albright & Hayes, 2003; Bybee, 1995, 2001; Pierrehumbert, 2003a, 2003b), and it tests whether one or both levels of variability are necessary for generalizing a phonotactic sequence to a new word.

This test of abstraction is relevant to three classes of models—exemplar models, exemplars-plus-abstractions or hybrid models, and symbol-manipulating models. Each model has been used to account for phonotactic frequency effects, but their underlying mechanisms differ greatly. Most critically, the models diverge with respect to the presence of abstraction. At one extreme, exemplar models are essentially abstraction free. At the other extreme, symbol-manipulating models presuppose certain types of abstractions. In the middle are models whose abstract means for abstract learning piggybacks on exemplar models. The results from the experiments described below provide a test of each model's means of representing phonotactic learning, present challenges for each model, and ultimately provide a set of data to which future theoretical work may be held accountable.

1.4 The Structure of the Dissertation

The dissertation is broken up into seven chapters. Following this introduction, Chapter 2 provides a review of the literature on phonotactic effects in language acquisition. This is followed by a review of the relevance of the perceptual learning mechanism as it relates to phonotactic learning. Then the three classes of models mentioned above are examined in more detail, and each is scrutinized and analyzed to make predictions in an artificial language learning paradigm. The predictions relate to three experimental manipulations: generalization of a phonotactic sequence from token variation alone, generalization from the combination of type frequency and token variation, and generalization from type frequency alone.

Chapter 3 reports an experiment that tests for generalization from token variation. Children listened to multiple talkers produce a word containing a target phonotactic sequence and were subsequently asked to produce a related word. Chapter 4 reports the results from an experiment testing for generalization from type frequency and token variation. Children heard multiple words containing a target phonotactic sequence and multiple talkers producing each word. Chapter 5 is a replication of the 'generalization from type frequency and token variation' experiment using an additional set of words. Chapter 6 uses the same word set from the previous experiment, but removes token variability: each word associated with a target phonotactic sequence was spoken by a single talker.

Chapter 7 provides an an omnibus analysis of all four experiments, as well as detailed comparisons of the results of each experiment with the others. The chapter includes an analysis with the separate experiments as conditions, a discussion of a possible ceiling effect in all four experiments and an analysis of the relative contributions of perceptual learning and repeated articulations to the accuracy measurement. I also review evidence for a syllable structure effect related to the English Frequency factor. Finally, three proposals for future research are laid out, including an additional experiment with 12 talkers, a study looking at the nature of word "types", and a study of the relationship between perceptual learning and sonority.

Chapter 8 is a general discussion of the four reported experiments. I review the motivation for the four experiments, the results of those experiments, and the interpretation of those results with respect to representations and mechanisms that have been called on to explain phonotactic probabilities. The findings are framed within the scientific investigation of phonotactic learning, language acquisition, and linguistics, in general.

CHAPTER 2

PHONOTACTIC PROBABILITY IN LANGUAGE ACQUISITION

In this chapter, I present a discussion of past research on phonotactic learning and a theoretical justification for the experiments discussed in subsequent chapters. Section 2.1 begins with a discussion of findings from the infant perception literature that suggest that knowledge of phonotactic probabilities can be linked to word learning. Similar findings in babbling and child speech production are also reviewed. Section 2.2 describes a series of studies suggesting how perceptual learning can link the findings from the infant and child language literature, and Section 2.3 describes how type frequency has been implicated as the critical factor for generalizing perceptual knowledge. The final section, 2.4.1, begins with a short description of the four experiments to be conducted. The experiments are placed in the context of three models of phonotactic learning, and predictions are made for the experiments based on the properties of each model.

2.1 Frequency Effects in Two Areas of Development

The influence of the ambient language on a child's linguistic development has been documented in a variety of experiments. Generally, these studies show a broad range of frequency effects, including the frequency or probability of a caregiver's voice, the predominant stress pattern of a language, and phonotactics. These frequency effects have been found using a variety of measurements, including the sound preferences of infants, the phonetic characteristics of babbling, and in children's earliest speech. I divide this research into two classes of studies: frequency effects in perception and frequency effects in production. The perceptual studies have largely been conducted with infants, the productions studies with older children. Although the pairing of ages with processing domains is largely the consequence of developmental limitations (production studies are not possible with infants), it has also created some confusion about the relevance of one domain to the other. This problem has at least a partial solution in the perceptual learning literature.

2.1.1 Frequency Effects in Perception

In an early study of ambient language influence, Mehler, Bertoncini, Barrière, and Jassik-Gerschenfeld (1978) showed that infants prefer their mother's voice to the voices of other women. Mehler et al. (1988) followed up on this finding by showing that newborn infants also preferred speech from their mother's native language to speech from another language. These studies provide evidence that infants have amassed some knowledge of their ambient language within days of birth. That conclusion is strengthened by DeCasper and Fifer (1980), who suggested further that infants' native language preference is the result of prenatal exposure to their mother's voice. Taken as a whole, these studies of newborn linguistic development confirm the importance of learning in the earliest stages of language acquisition.

As infants grow older, they are able to further delineate native language patterns (Werker, Gilbert, Humphrey, & Tees, 1984; Werker & Tees, 2002) and begin to attune to the phonological patterns that make up their native language's phonological structure. One such pattern is the predominant stress pattern of the language. Jusczyk, Cutler, and Redanz (1993) used the head-turn preference procedure developed by Fernald (1985); Kemler Nelson et al. (1995) to show that, by 9 months of age, English learners prefer a strong-weak, or trochaic, pattern of syllables over a weak-strong, or iambic, pattern. The authors attributed this preference to the influence of the ambient language: trochees are more common in English than iambs.

In a study of infant's knowledge of phonotactics, Jusczyk, Friederici, et al. (1993) compared English and Dutch infants' preferences for native and non-native phoneme sequences. The phonemic inventories of Dutch and English include /k, d, n, w/, but the languages differ as to how these phonemes can be combined in syllable onsets. Dutch allows /kn/ onsets, as in *knevel*, but not /dw/ onsets, as in the English word *dwindle*. The opposite pattern (/dw/ is licit, /kn/ is not) is true of English. Jusczyk et al. found that 9 month-old Dutch and American infants preferred native language phonotactic sequences over non-native sequences. The results, combined with findings from Mehler et al. (1988) and Jusczyk, Cutler, and Redanz (1993), show that infants are sensitive to patterns in the ambient language by a very early age. Additionally, they suggest that language acquisition may be guided by, may possibly even be dependent on, native language experience.

Jusczyk et al. (1994) made an important extension of this study by addressing infants' sensitivity to phonotactic probabilities, not simply native versus non-native phonotactics. Rather than comparing infants' sensitivity to phonotactic sequences that were either present or absent in English, Jusczyk et al. had infants listen to nonsense words with high or low phonotactic frequency. Although all of the nonsense words in this study were permissible according to categorical English phonotactics, one list contained words like /t͡ʃʌn/[1], with common, or high probability, phonotactic sequences, and the other list contained words like /ɡiθ/, with less common, or low probability, phonotactic sequences. The authors found that children preferred the words with high probability phonotactic sequences, suggesting that children are attuned not simply to general language phonotactics, but to the relative frequencies of the phonotactic patterns to which they are exposed. Furthermore, the results reinforce the position that phonological development is guided by patterns in the input.

In a more recent study, Chambers et al. (2003) showed that infants could learn phonotactic patterns in an artificial language learning paradigm after a four minute exposure. 16.5-month-old infants were familiarized with sets of words in which one set of phonemes was consistently assigned as onsets and the other set as codas. Half of the infants were familiarized with a list that used the sounds /b, k, m, t, f/ as onsets and /p, ɡ, n, tʃ, s/ as codas. Test words in this list included /bɪp/ and /mɪn/. The other half heard a list in which the sounds were assigned in the reverse order (/pɪb/ and /nɪm/). Using the head-turn preference procedure, Chambers et al. tested the infants to see if they learned the onset and coda patterns. In the test, infants listened longer to lists where the onset and coda assignments were reversed, suggesting that they had learned the phonotactic

[1] Throughout this dissertation nonsense words will be transcribed using the International Phonetic Alphabet (International Phonetic Association, 1999) rather than an approximated spelling

patterns. By exposing infants to multiple words that share phonotactic properties, Chambers et al. showed how phonotactics can be learned rapidly from a representative word set. Like the studies conducted by Jusczkyk and colleagues, the results from this study suggest that language experience guides phonological development. Infants store words in memory and make generalizations about those stored forms. A similar but somewhat more limited set of findings for 9-month-olds are discussed by Saffran and Thiessen (2003).

Another set of studies has focused on how phonotactic probabilities relate to word learning. Hollich, Jusczyk, and Luce (2002) looked at how familiarization with multiple examples of a phonotactic sequence influenced an infants' ability to create form-referent pairings. The authors used a combination of the head-turn preference procedure and the split-screen preferential looking paradigm to familiarize 17 month-old infants with the target phonotactic patterns and test how well infants associated a word with the target patterns to a referent. In the familiarization, infants heard words like /tɚŋ/ and /tɚtʃ/, which both contain the CV sequence /tɚ/, and /θɚb/ and /mɚb/, which contain the VC sequence /ɚb/, all of which are related to the test word /tɚb/. The infants in this study were better at linking /tɚb/ to a referent when they had been familiarized with a larger number of phonotactically related words, suggesting that known phonotactic patterns can facilitate word learning.

The infant perception literature makes a strong case for the role of ambient language frequency patterns in phonological acquisition. More specifically, the infant literature attributes the development of phonotactic knowledge to an infant's ability to learn the relative frequency of ambient language patterns through perception. Infants appear to learn sub-lexical patterns like phonotactics by storing detailed perceptual word-forms in a receptive lexicon[2], over which computations about the probability of particular patterns can be made. In the next section, studies of child speech production are reviewed. The results of these studies reveal similar effects of ambient language frequency, including

[2] In the literature, *receptive lexicon* refers to the set of words which children can understand. I am using it here to mean the set of words that children have learned through perception, but that may not have any referent. No unique phrase has been developed to describe the latter type of knowledge, but the use of *receptive lexicon*, at least in Edwards et al. (2004) appears similar to its use here. An alternative term could be *word-form lexicon* or *lexeme store*.

in studies of phonotactics, strengthening the claim that phonotactic probabilities are a consequence of word learning.

2.1.2 Frequency Effects in Production

Babbling

Ambient language effects are not limited to perception. Some of the first evidence of such frequency effects in linguistic development came from the careful study of the phonetic properties of babbling. Boysson-Bardies, Sagart, and Durand (1984) made acoustic recordings of the babbling of babies learning four different languages (English, French, Japanese, and Swedish). They played their recordings to adult speakers of those languages. They found that the adults could discern which babbles came from babies learning their native language and which babbles did not. Boysson-Bardies et al. concluded that, even before words begin to emerge, experience with the ambient language has begun to mold early vocalizations.

In a subsequent study, Boysson-Bardies, Hallé, Sagart, and Durand (1989) performed a cross-linguistic study of the vowels in babbling. They transcribed vowels and measured vowel formants for infants learning Arabic, Chinese, English, and French. Although all infants tended to produce vowels in three categories—low front, mid central, and low central—the transcription analysis showed that the four groups produced different quantities of each category. English learning infants, for example, produced more front vowels, whereas French learning infants produced more mid central vowels, consistent with the frequency of those vowels in the native language. Boysson-Bardies et al. also found consistent differences in the formant measurements for each group of infants that reflected average formant values of the infants' native language. The authors concluded that babbling shows evidence of the child's attempt to incorporate knowledge of their native language into their own productions. Before children match word-forms to semantic content, the phonetic qualities of their productions match the qualities of adult models.

In another study of babbling, Boysson-Bardies and Vihman (1991) analyzed the labial consonants produced during babbles and early speech by 9-17 month-old infants learning English, French, Japanese, and Swedish. They found that the occurrence of labials was significantly higher for infants learning Swedish and French, a difference that corresponded to the relatively higher frequency of labials in those two languages. From these experiments, it is clear that the influence of the native language appears in a child's earliest productions.

Early Words

There is evidence that the frequency of particular linguistic units, such as phonemes, has an effect on the production of those units. Boysson-Bardies and Vihman (1991) showed that the first 25 words in a child's expressive vocabulary tend to contain the most common phonemes of the child's native language. For example, labials such as /w/ have a moderately high frequency in English, a higher frequency in French, but are relatively low frequency in Japanese and Swedish. Boysson-Bardies and Vihman showed that children's use of different consonant classes corresponds to the frequency of those classes in the language, so children acquiring French use the most labials, followed by children learning English, and the fewest number of labials were used by children learning Japanese and Swedish.

Effects of phonotactic frequency specificially have also been observed in early words. A study by Messer (1967) showed that children were able to produce permissible phonotactic sequences in a new word, but struggled if the word contained an unknown phonotactic sequence. Preschool aged children (3;1 to 4;5) were presented a pair of nonsense words. One member of the pair contained an onset conforming to English phonotactics, such as /frul/; the other member did not, such as /mrul/. Children were asked to repeat back the word that sounded most like an English word. Messer found that children preferred the words with English phonotactics and they produced those words more accurately. The results confirm that children are sensitive to the phonotactic permissibility of a sound sequence and are most accurate when producing licit phonotactic sequences.

D. K. Ohala (1996, 1999) showed that production accuracy also depends on the phonetic or featural makeup of a phonotactic sequence. Two-year-old children were asked to produce nonsense words containing unattested onsets such as /bw/, as in the word

/bwɪv/, and /tm/ , as in the word /tmaʊd/. Neither /bw/ nor /tm/ occur as onsets in English, but the former is similar to the stop-glide clusters /tw/, /dw/, /kw/, and /gw/, which do appear in the language, whereas the latter represents an unattested stop-nasal cluster. Children were more accurate when producing clusters like /bw/ compared to clusters like /tm/, suggesting that children's productions are not only sensitive to phonotactics, but also to a phonotactic sequence's internal structure.

Beckman and Edwards (1999) showed that children's productions are influenced by probabilistic phonotactics. The authors conducted two experiments in which children between 3;2 and 5;0 years of age were asked to repeat nonsense words with target phonotactic patterns of varying frequency. The words were composed of two or three syllables, and each contained two target sequences that were either common, or high probability, in English (e.g., the biphone /sn/ in /busnədi/) or uncommon (/bn/ in subnədi/. All words were matched on a wordlikeliness rating (cf. Gathercole, Willis, Emslie, & Baddeley, 1991) and were presented to the children via computer. Children were asked to repeat the words immediately after presentation. Beckman and Edwards found that children were significantly more accurate at producing the high probability sequences. The second experiment also included children with phonological disorders. Although disordered children were less accurate overall, they, too, produced high probability sequences more accurately than low probability sequences. Beckman and Edwards concluded that productive phonology must rely on phonotactic knowledge derived from stored lexical forms because accurate speech production for both normally and atypically developing children depends on the frequency with which phonotactic patterns appear in their native language.

Similar results were found by Munson (2001). Using a number of measures of speech production, including accuracy, speed, and fluency, he looked to see how each measure was affected by a manipulation of the phonotactic probability of a medial consonant sequence, or cluster. Two groups of children (younger group's mean age was 3;10, older group's was 8;4) produced high frequency word-medial consonant clusters (e.g., /st/ in the nonsense word /nɪstəp/) faster, more accurately, and with less variability in duration than low frequency clusters (e.g., /fp/ in the nonsense word /nɪfpət/). Munson concluded that phonotactic knowledge is relevant to various components of the speech production mechanism, and that the findings across a variety of measures suggest that phonotactic probabilities are encoded at multiple levels, from abstract phonological knowledge to articulatory timing.

In a study of the relationship between phonotactic probability and the lexicon, Edwards, Beckman, and Munson (2004) looked at children's (104 children between 3;2 and 8;10) production accuracy for nonsense words varying in phonotactic probability. They correlated children's accuracy scores with scores on several standardized vocabulary tests, that is, measures of the children's lexicon size[3]. In addition to finding that sequences with higher phonotactic probabilities were produced more accurately than sequences with lower probabilities, they found that vocabulary size correlated with production accuracy. When age and intelligence were held constant, children with larger vocabularies were more accurate in their productions than children with smaller vocabularies. Edwards *et al.* linked this finding to Pierrehumbert's (2001, 2003a, 2003b) theory of phonological acquisition, which proposes that phonological knowledge is the accumulation of generalizations over the lexicon. More will be said about this theory in the discussion of bottom-up and top-down models of phonological development in Section 2.4.1.

Another study which convincingly shows the importance of phonotactic probability in word production was conducted by Zamuner et al. (2004). They were concerned with whether the production of a single phoneme might be affected by the frequency of its environment. To address this possibility, they created CVC nonsense words with constant neighborhood density and word-likelihood ratings, and they varied the phoneme and diphone frequency of the coda consonants. For example, /l/ is more likely to appear in an environment like that of /gɛl/ than that of /pʌl/. Children between 1;8 and 2;4 years of age produced the consonants in high-frequency phonotactic environments more accurately than the same consonants in low-frequency environments, suggesting that their productions were mediated by the frequency of the phonotactic environment.

The studies discussed above suggest that word learning guides phonotactic learning,

[3] the term *vocabulary* is used here to refer to a measure of the number of words that children have stored in memory. The term *lexicon* refers to the psychological or theoretical collection of stored words.

but the opposite is true, as well. Storkel and colleagues (Storkel, 2001, 2004; Storkel & Morrisette, 2002; Storkel & Rogers, 2000) have conducted a number of studies documenting the effects of phonotactic frequency on word learning in older children. Storkel and Rogers (2000) showed that children of 10 and 13 years of age are better at pairing novel words to objects when those words have common sound patterns. Although Storkel and Rogers did not find an influence of phonotactics for 7 year-olds, a subsequent study (Storkel, 2001) showed phonotactic effects in noun learning for children as young as 3;2. In that study Storkel (2001) looked at children between the ages of 3;2 and 6;3. Generally speaking, children learned to associate the phonetic word forms to referents faster when the word forms were composed of common phonotactic sequences. Difficulty with words composed of infrequent phonotactic sequences, Storkel suggests, is the result of increased processing demands to handle the rare sound sequences, resulting in weaker semantic representations (cf. work on phoneme selection and avoidance, e.g., Schwartz & Leonard, 1982). Similarly, Storkel (2003) found facilitative effects of phonotactic probability in a study of verb learning, suggesting a broader role for phonotactic frequency in word learning than had been shown previously (but cf. Storkel (2004) for less positive results for children with functional phonological delays and Maekawa and Storkel (2006) for individual differences).

Like the infant perception literature, studies of child speech production provide overwhelming evidence that phonotactic probabilities reflect lexical statistics (cf. Edwards & Beckman, 2008 for related cross-linguistic evidence), rather than grammatical/ungrammatical rules, even in the expressive lexicon[4]. Given the overlapping findings in the processing domains of perception and production, it is natural to consider the possibility that a common mechanism underlies both the perceptual knowledge of infants and the production differentials of young children. Until recently, however, it has been unclear whether one mechanism could be responsible for creating representations linking both the perception and production systems.

[4] Unlike *receptive lexicon*, there is no clear connotation for semanticity in the term *expressive lexicon*. This fact further justifies the present use of the two terms.

2.2 Linking Perception and Production

2.2.1 Perceptual learning as it influences speech production

There is an empirical divide separating the infant and child literatures. Although both areas of research detail effects of phonotactic probability on speech processing, the former is limited to perceptual processing, and the latter is concerned primarily with speech production. This divide is somewhat artificial in that it is largely due to the fact that infants are not able to speak. However, many accounts of speech development suggest that the perceptual learning seen in infant studies is relevant to speech development. According to this *perceptual learning* hypothesis, representations accrued through perception are stored in such a way that they facilitate the production of those same patterns. In fact, perceptual learning has been shown to affect the results of both perceptual and production-based tasks. It has been implicated in measures from both domains of processing, and several studies have concluded that perceptual learning can be used to explain perceptual- and production-based frequency effects. This research includes a study by Gerken et al. (2006), discussed in Section 2.2.2 below, that inspired the experiments conducted for this dissertation.

In this dissertation, I adopt the perceptual learning hypothesis. There are two important reasons why I am focusing exclusively on perceptual learning as an explanation for phonotactic probability effects in production. First, a sizable literature has amassed supporting claims for a perceptual learning mechanism. These studies, which are discussed below, have shown very clear facilitative effects of perceptual learning on speech production. Second, given the fact that perceptual learning is already implicated as an important feedback mechanism supporting speech development, there is very little support for purely articulatory hypotheses of speech development.

The choice of the perceptual learning hypothesis means that I am leaving aside alternative explanations for how speech develops. One such proposal is that repeated articulations, or *articulatory practice* , is the mechanism behind phonotactic effects in speech production. There is some evidence that articulatory practice plays an important role in speech development (Edwards et al., 2004; Messum, 2007), and there are cases where children often have accurate perceptual representations of a sound and nevertheless make inaccurate articulations, as is the case in covert contrasts (cf. Gibbon, 1999). It is also possible that articulatory practice and perceptual learning contribute to shared representations, as is suggested by MacKay's (1989) Node Structure Theory, or that articulatory practice makes an independent contribution to speech development, but I leave those possibilities aside. For the present, it is sufficient that the perceptual learning hypothesis is well motivated as a mechanism for improving speech production. This allows us to move beyond basic questions about the existence of the mechanism and explore what kinds of perceptual information children use to make phonotactic generalizations. I turn now to the evidence for the perceptual learning mechanism.

2.2.2 Research on Perceptual Learning as it Affects Production

The idea that perception facilitates or even enables production has been well developed in the second language acquisition literature. Contrary to claims that no correlation exists between how L2 learners perceive and produce the nonnative language, Flege's (1995) Speech Learning Model predicts nearly the opposite: that a correlation does exist and accuracy in L2 production is largely limited by L2 perception accuracy. His claim is bolstered by numerous empirical studies of the limited production capabilities of L2 learners. Flege (1993), for example, showed a correlation between how Chinese learners of English perceived and produced the vowel length distinction that signals the voicing of word-final consonants. In a large study of L2 vowels, Flege, Bohn, and Jang (1997) looked at the perception-production relation for German, Spanish, Korean, and Mandarin learners of English. Generally, Flege et al. found that learners who relied on spectral cues rather than durational cues to perceive tense-lax and height distinctions (e.g. /i/-/ɪ/ and /ɛ/-/æ/) were also more likely to produce vowels with spectral distinctions. Flege (1999) provides a more substantial review of the evidence that second language (L2) productions are dependent on L2 perceptual accuracy.

Additional support for the claim that perceptual accuracy places limitations on production comes from work by Lotto, Sato, and Diehl (2004). They compared the productions of /r/ and /l/ by native English speakers to Japanese speakers' productions of the

Japanese flap, /ɾ/, and American /r/ and /l/. The /r/-/l/ contrast is notoriously difficult for native Japanese speakers, and Lotto et al. were interested in whether the acoustic qualities of existing Japanese contrasts contributed to that difficulty. Both the English and Japanese speakers produced six two-syllable words of English, (reading, rocking, rooting; leading, locking, looting); the Japanese speakers produced Japanese words beginning with /ɾ/, as well. As expected, native English speakers largely separated their /r/ and /l/ productions acoustically using the third formant (F3), and the Japanese /ɾ/ was distributed across the English category boundary. The productions of English words by Japanese speakers were also separated, but largely by F2 rather than F3. This follows from the usefulness of F2 in separating /ɾ/ from /w/ in Japanese, but also from the fact that F2 and F3 are highly correlated in /ɾ/ productions. Lotto et al. conclude that Japanese speakers must be using native-language-based acoustic targets (points on F2) to when producing the non-native contrast, even though this perceptual strategy is detrimental to L2 communication.

Perception limits L2 acquisition, but perceptual learning has also been shown to facilitate the acquisition of a non-native contrast. Work by Anne Bradlow and colleagues (Bradlow et al., 1997, 1999) has shown that perceptual training facilitiates L2 speech production. In Bradlow et al. (1997), native Japanese listeners completed several weeks of perceptual training on the /r/-/l/ contrast. Participants completed a series of word identification tasks that used a set of English minimal pairs contrasting the two sounds in different phonological environments and spoken by different talkers (Logan, Lively, & Pisoni, 1991). Participants also completed pre- and post-tests in which they were asked to produce the set of training words and a set of novel words. Because the latter set consisted of unencountered words, participants needed to possess generalized knowledge about the /r/-/l/ contrast to make accurate productions. To evaluate the efficacy of the training, Bradlow et al. played the Japanese speakers' pre-test and post-test productions to native English speakers in an intelligibilty rating task and a forced-choice identification task. The results from both tasks showed that the perceptual training resulted in improved speech accuracy compared to pretest productions and compared to a control group that did not receive the training.

In Bradlow et al. (1999), the authors brought both the training and control groups back following a 3-month delay to see whether the facilitative effects of the perceptual training were maintained long-term. Subjects performed the same posttest production task, and English speakers were asked to evaluate their productions in the same intelligibility rating and forced-choice identification tasks, as well as an additional transcription task. The results from all three measures showed that the Japanese speakers had maintained a level of production accuracy over the three-month delay that was significantly better than their pretest productions and the productions of the control group.

Wang et al. (2003) present evidence that perceptual training can have lasting benefits on the production of nonnative tonal contrasts. They report data from an experiment where native English speakers with limited experience with Mandarin completed a perceptual training on the four Mandarin tones as spoken by multiple Chinese speakers in a number of different words. Native Mandarin speakers then compared L2 learners' productions of Mandarin tones before and after the training to a control group with no training. Wang et al. found an interaction between ratings of pre-training and post-training productions and the different subject groups: the perceptual training group showed significant and long-lasting effects of the training (or a significant difference between pre-training and post-training) but no effect was found for the control group. Furthermore, the benefits of perceptual training generalized to productions of the four tones in words not included in the training.

The L2 literature provides strong support for the claim that speech development is dependent on perceptual learning. Studies of diverse populations learning a number of different languages have shown that perceptual training can improve speech production. Importantly, the perceptual training regimens used in these studies share certain characteristics. L2 learners were exposed to a variety of words spoken by a variety of talkers in a high variability perceptual input (Logan et al., 1991). In the next section, I discuss experiments showing how important perceptual variability can be to child speech.

The 10 Talkers Study

Recently, the effects of perceptual training on speech production have also been extended to L1 learning. The basic question asked by Gerken et al. (2006) (referred to hereafter as the 10 Talkers study) was whether manipulating the number of times a word was heard, or the word's *token frequency*, influenced a child's ability to produce that word. In the tradition of artificial language learning experiments like Chambers et al. (2003), Gerken et al. wanted to create perceptual frequency effects within the experiment. The assumption was that a high enough level of experimental frequency should result in perceptual learning and improved speech production speed and accuracy.

To test this possibility, Gerken et al. (2006) familiarized children with a set of nonsense words and varied the number of times children heard each word. Four-year-old children were familiarized with CVCCVC nonsense words that were described as make-believe animal names (D. K. Ohala, 1999). The words were created in pairs. Pairs of nonsense words were identical except for the medial consonant sequences (hereafter consonant clusters), which varied according to their English Frequency. The word set, broken down by the High English (hereafter HiEng) and Low English (hereafter LoEng) conditions, is given in Table 2.1. The word pair /fæmpɪm/ and /fæmkɪm/, for instance, are identical except for the medial cluster. The /mp/ in /fæmpɪm/ is a common English cluster, whereas the /mk/ cluster in /fæmkɪm/ is quite rare.

Table 2.1: Words used in the 10 Talkers study (Gerken et al., 2006). The words varied in terms of the English Frequency of the medial consonant cluster, but word pairs were matched for their surrounding environment.

High English	**Low English**
/fospəm/	/foʃpəm/
/mæstəm/	/mæfpəm/
/fæmpɪm/	/fæmkɪm/
/boktəm/	/bopkəm/

A word's Experimental Frequency was a manipulation of the number of times the word was heard during the familiarization, or the word's *token frequency*. A given word was presented either 10 times (Experimental High condition, hereafter ExpHi) or only once (Experimental Low, hereafter ExpLo); each word served as both ExpHi and ExpLo in the experiment, although a given child heard each word in only one of the two Experimental Frequency conditions. The Experimental Frequency factor was designed to create artificial probabilities. The expectation was that words in the ExpHi condition would be produced more accurately than words in the ExpLo condition, similar to the finding that EngHi clusters are produced more accurately than EngLo clusters.

Following the familiarization, children repeated the words in a repetition task. In the first experiment, token frequency was simply the total number of times that a word was played, that is, each token of an ExpHi word was identical. Gerken et al. (2006) reasoned that, if a simplistic view of the perceptual frequency hypothesis is correct, children should show improved accuracy for the words they heard frequently. Results from this experiment are given in Figure 2.1. As expected based on previous research on phonotactic probabilities in child speech (Beckman & Edwards, 1999; Edwards et al., 2004; Munson, 2001; Zamuner et al., 2004), children were more accurate when producing the words with HiEng clusters (blue and white bars clustered on the left in the upper half of Figure 2.1) compared to the words with LoEng clusters (blue and white bars clustered on the right). Although there is a trend towards more accurate productions of the ExpHi words (blue bars) compared to the ExpLo words (white bars), children were not significantly more accurate as a result of the Experimental Frequency manipulation. No effect of either English Frequency or Experimental Frequency was found in the analysis of production latencies (lower graph in Figure 2.1). This result suggests a limited role for token frequency in perceptual learning.

Although the results of the first experiment appear counter to the perceptual learning hypothesis, Gerken et al. note that perceptual learning may not depend solely on raw token frequency. Perceptual learning may instead rely on properties of speech which are known to affect speech perception (Logan et al., 1991). Such properties include talker identity, which can vary even when the linguistic message is constant. Throughout the dissertation, I will refer to this property of the input as *token variability* or *phonetic variability*, which refers to the fact that a given word may vary along a number of phonetic dimensions without resulting in a change of meaning. Talkers have different vocal tract lengths and vocal band widths, affecting the relative values of their formants and their fundamental frequency (pitch), respectively. They may speak fast or slow, and no two

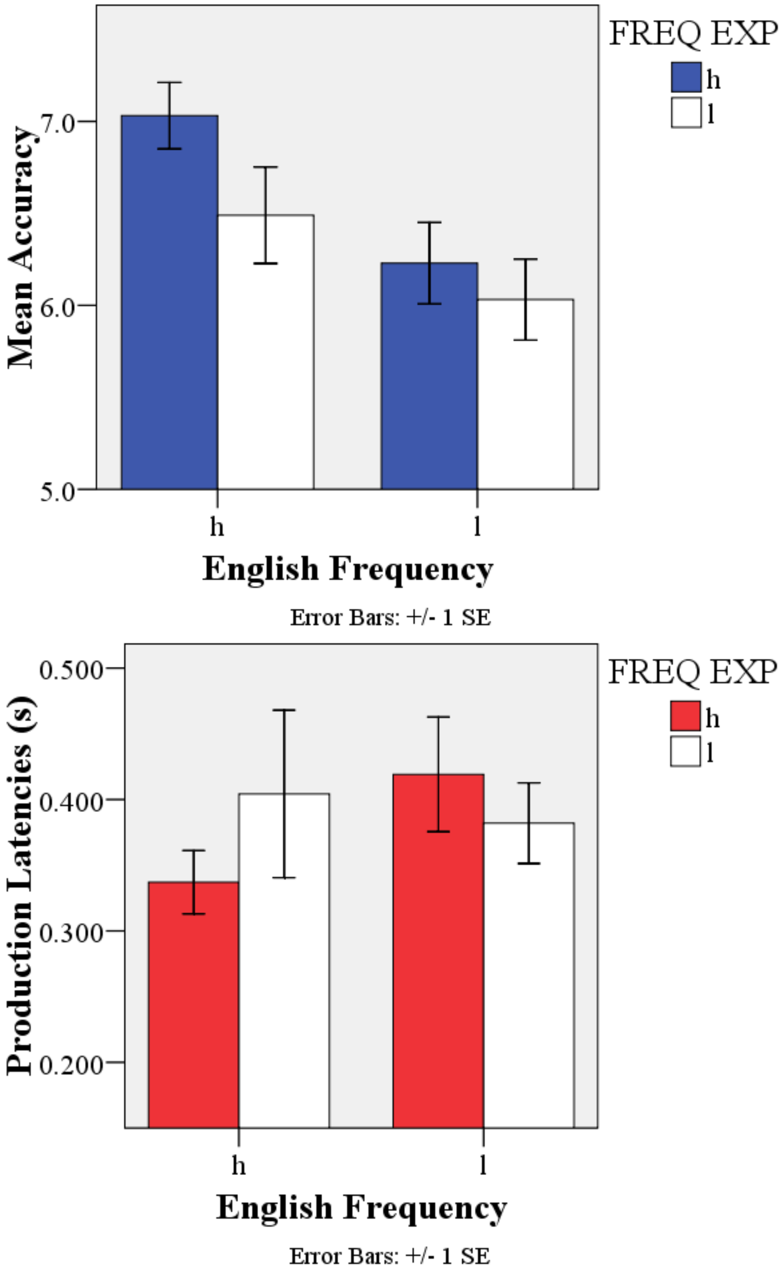


Figure 2.1: A reproduction of the bar graph plots of the accuracy (top) and production latency (bottom) analyses of the first 10 Talkers experiment, in which experimental High words were presented 10 times, with the same acoustic token each time. No significant effect of Experimental Frequency is seen in either dependent measure.

talkers pronounce things in exactly the same way; yet none of this variability presents a serious challenge for adult listeners.

One reason that phonetic variability may not impede perceptual learning is that listeners store these details, sometimes referred to as indexical information, and they may use token variability to facilitate subsequent speech recognition. For example, Goldinger (1996) showed that adults have an impressive capacity to remember talker-specific information, and memories for talker-specific phonetic qualities can facilitate word recognition even a week after the voice was first introduced. This finding suggests that people are well equipped to both perceive and store phonetically variable tokens. Houston and Jusczyk (2003) showed that infants are better at learning a word and recognizing it in fluent speech if the infant was familiarized with multiple speakers producing the word. Similarly, Singh (2008) found that the variability present in different affective qualities produced by the same talker also facilitates infants' word recognition. In combination, these results suggest that exposure to some form of phonetic variability is not simply harmless, it may actually be a critical component to perceptual learning. This adds detail to the perceptual learning hypothesis because it suggests that hearing something repeatedly is not enough to facilitate learning. Auditory exposure must also include some level of variability so that learners can identify which dimensions of the stimuli are most relevant (cf. Holt & Lotto, 2006 for additional discussion).

To address the potential role of phonetic variability in perceptual learning, Gerken et al. added token variation to the familiarization phase of a second experiment. The design was identical to the first experiment, except that children heard words in the ExpHi condition produced by 10 different talkers. Results are given in Figure 2.2. This time children were significantly more accurate (blue bars in upper graph in Figure 2.2) and faster (red bars in lower graph) as a result of the token variation manipulation. In fact, the strength of the Experimental Frequency factor appears to have washed out the expected English Frequency factor. The results provide strong support for the perceptual learning hypothesis and for the role of perceptual variability in facilitating word learning (Houston & Jusczyk, 2003; Singh, 2008).

Comparing the results of the first and second experiments, the 10 talkers study sug-

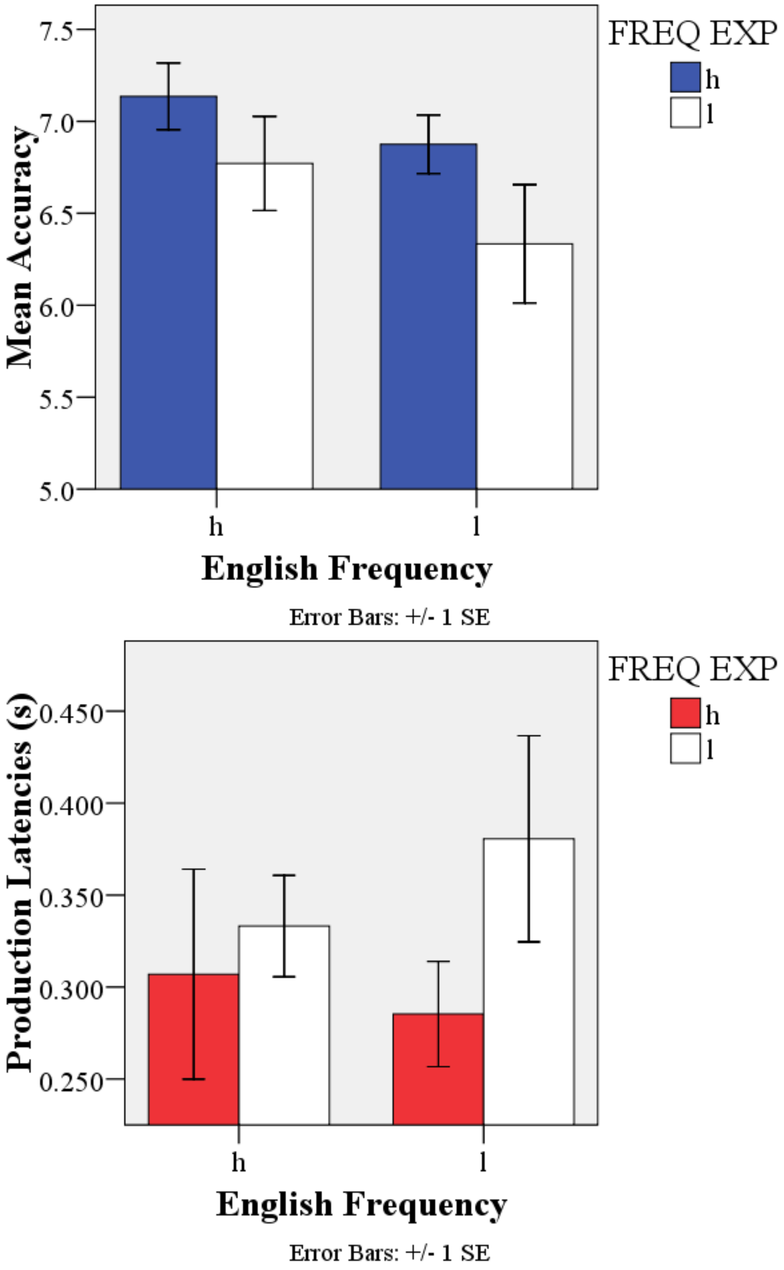


Figure 2.2: A reproduction of the bar graph plots of the accuracy (top) and production latency (bottom) analyses of the second 10 Talkers experiment, in which experimental High words were presented 10 times, with the different talkers producing each token. Children's productions for the Experimental High words were faster and more accurate compared to their productions of the Experimental Low words.

gests that the same facilitative effects of perceptual learning that are seen in the L2 literature (Bradlow et al., 1997, 1999; Flege, 1999; Wang et al., 2003) are relevant in L1 acquisition, as well. Children produced novel words faster and more accurately when they were first exposed to multiple talkers producing that word. It appears, then, that perceptual learning is an important component to learning to produce a first language.

The 10 Talkers study also provides a clearer picture of what perceptual learning must entail. The simple token frequency manipulation in the first experiment was not particularly effective in changing productions, whereas the token variability manipulation in the second experiment provided robust facilitative effects. It follows that phonetic variability, such as that present in the voices of multiple talkers, is an important factor in perceptual learning.

In addition to supporting the perceptual learning hypothesis, the 10 Talkers study provides a potential explanation for phonotactic learning. Consider, for example, the results reported by Edwards et al. (2004). It is possible that the correlation between accuracy in production and lexicon size that those authors found reflects perceptually-based word learning, at least in part. This is further supported by the fact that perceptual learning facilitated productions of words containing both high probability (HiEng) and low probability (LoEng) phonotactic sequences. However, to conclude definitively that perceptual learning is behind the phonotactic probability effects seen in children's productions in the Edwards et al. study and in other production studies of phonotactic probabilities (Beckman & Edwards, 1999; Munson, 2001; Zamuner et al., 2004), additional evidence is necessary. In the 10 Talkers study, children were familiarized with and asked to produce the same words. This design does not allow us to determine whether children made any kind of word-internal analysis of the experimental words, or whether the facilitative effects of token variability affected anything other than word learning. In the 10 Talkers study, word learning and phonotactic learning are inseparable. For evidence of phonotactic learning, it must be shown that perceptual learning of a phonotactic sequence, separable from the word or words it appears in during familiarization, will cause a change in speech production. Children must learn about a phonotactic sequence apart from its presence in any given word, and then apply this perceptual knowledge to the production of a new word with the same phonotactic properties. That is, children must *generalize* a perceptually trained phonotactic sequence to their own speech.

The L2 literature provides some evidence that people can and do generalize phonological patterns learned in perception and applied to speech. Bradlow et al. (1997, 1999), for example, showed that Japanese learners of English generalize perceptually trained /r/ and /l/ categories to new words. Wang et al. (2003) showed that American learners of Mandarin are more accurate at producing tone contrasts in new words following a perceptual training. Several important questions remain unanswered by this research, however, with respect to the present focus on phonotactic learning. First, it is unclear whether L1 learners will behave similarly, because studies involving generalization based on perceptual learning are confined to the study of second languages. Second, neither the Bradlow et al. nor the Wang et al. studies were about phonotactics. Rather, they were about learning nonnative contrasts. Third, all of the L2 experiments used a combination of factors in their perceptual training, including multiple training words, multiple talkers, and multiple phonotactic environments. Therefore, we cannot determine from these studies which factor or combination of factors is most important for promoting generalization.

To better understand which factors of the ambient language might be most important for generalization, we can turn to work in theoretical linguistics. Broadly speaking, linguistic studies of generalization are concerned with phonological induction, or the hypothesis that phonological patterns are learned over the set of words that share a common phonological property. This research has focused on learning morphological correspondences, such as the rule for past tense formation in English, and is still ongoing (cf. Newman, Ullman, Pancheva, Waligura, & Neville, 2007), but most experts agree that generalization based on word-type frequency, or the number of unique lexical items sharing a pattern, plays a critical role in learning at least some types of morphological correspondences, if not all correspondences[5]. Furthermore, scientists on both sides of the association/symbolic processing debate would agree that the gradient nature of phonotactic probabilities is best handled in an inductive system. In the next section I review the

[5] See Pinker and Ullman (2002) and surrounding debate, conveniently available on Michael Ullman's website: http://brainlang.georgetown.edu/publications_by_year.htm

morphological correspondence literature which supports the role of type-based induction or learning. I show how these studies of the relationship between type frequency and morphophonology have led to current theories of phonotactic learning.

In a wide body of literature spanning studies of first and second language acquisition, research has linked perceptual learning to speech development. The core finding is that a variegated perceptual input enhances production accuracy. This was made especially clear in the 10 Talkers study (Gerken et al., 2006), in which children were familiarized with a set of words in perception and later asked to produce those words. The first experiment showed that simple perceptual frequency does not affect production. The second experiment added a level of talker variability to perceptual frequency, and this resulted in increased speed and accuracy in children's productions. Although the design of the 10 talkers study suggests that the effects of phonotactic probability seen in the child production literature (e.g., Edwards et al., 2004) may be attributable to perceptual learning, this conclusion cannot be drawn until perceptual learning is shown to generalize to new productions. The L2 literature suggests that this generalization is possible, but it leaves open questions about which qualities of a perceptual input are most important—do we learn phonotactics from a perceptually variable input, or by storing multiple related word-forms, or from the combination of the two? In the next section, I review theoretical research which points to the importance of multiple word forms, or type frequency, for generalization.

2.3 Type Frequency and Learning Morphological Correspondences

Type frequency refers to the total number of words in the lexicon that share a pattern. For example, the type frequency of the phonotactic sequence /bl/ is quite high, appearing in almost 500 words word-initially and thousands of words word-finally (counting b + syllabic l, as in the suffix *-able*, based on the MRC Psycholinguistic Database, http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm), whereas /sf/ has a low type frequency, appearing at the beginning of only about 15 words and in the middle of about 224 words. The importance of type frequency for phonological induction has been discussed at length by Bybee (1995, see also Bybee, 2001), who presented evidence that languages establish default morphological patterns, such as the phonological alternations that occur in the regular form of the English past tense *-ed*, based on high type frequency. Following similar claims made by connectionist modelers (e.g. MacWhinney & Leinbach, 1990; Plunkett & Marchman, 1991; Rumelhart & McClelland, 1986), Bybee claims that a phonological pattern can only hold to the extent that it is generally true of words in the lexicon. This claim is in contradiction to predictions of the dual processing model proposed by Pinker and colleagues (Clahsen & Rothweiler, 1992; Marcus et al., 1992, 1993; Pinker, 1991, 1999; Pinker & Prince, 1994; Prasada & Pinker, 1993), which posits a symbolic rule for regular morphology, regardless of the type frequency of words conforming to this pattern. Support for this model comes from the German system of past participles, whose regular form, according to Clahsen and Rothweiler (1992), has a relatively low type frequency. Bybee (1995) has disputed this claim, however, and shows how an alternative count of German participles results in the regular form having the highest type frequency. Additionally, she reviews a number of other cases where the regular pattern is clearly the highest in type frequency, including plural formation in German, Arabic, and Hausa.

Following Bybee's (1995) lead, subsequent studies of phonotactic probability have focused on how probabilities relate to type frequency. Bailey and Hahn (2001) showed that type statistics have important predictive power in explaining the differential effects of neighborhood density and phonotactic frequency on wordlikeness judgments. Rather than being concerned with whether type frequency or symbolic processing underlie abstract phonotactic knowledge, they questioned whether abstraction of phonotactics is necessary at all. They noted the correlation between phonotactic probability and neighborhood density, or the group of words which differ by only one phoneme, and designed an experiment to test whether lexical neighborhood activation (simultaneous activation of words starting with /tr/) could subsume phonotactic effects (activation of an abstract /tr/ entity). Their exemplar-based model of neighborhood activation, the Generalized Neighborhood Model, is discussed in greater detail in Section 2.4.1 below.

To determine whether phonotactic probability effects seen in wordlikeness judgments could be explained solely by neighborhood activation, Bailey and Hahn created nonsense word lists where neighborhood density and phonotactic probability were manipulated separately. For example, /slɪsk/ and /bɪnθ/ have virtually identical phonotactic scores, but /slɪsk/ has just two lexical neighbors (*slick* and *slink*, based on a search of the MRC database) whereas /bɪnθ/ has fourteen (e.g., *bent, bins, binge, built, dint, mint*, and *tint*). On the other hand, /prʌnθ/ and /ʃrʌpt/ have the same number of neighbors (no neighbors) but /prʌnθ/ has a higher phonotactic probability. Participants were asked to rate how much words like /prʌnθ/, /ʃrʌpt/, /slɪsk/, and /bɪnθ/ sounded like English words on a scale of 1 to 7[6]. The authors found that both neighborhood density and phonotactic frequency contributed to a nonsense word's wordlikeness rating. Bailey and Hahn conclude that type frequency does contribute to phonological learning. I will return to Bailey and Hahn's paper later, particularly to the exemplar model. It is the first of three models that were used to make predictions for a series of perceptual learning experiments.

Hay, Pierrehumbert, and Beckman (2003) also examined wordlikeness ratings and their relationship to type frequency, focusing on nasal-obstruent clusters, but they added a transcription analysis in which participants were asked to write down what they heard. Like Gerken et al. (2006), Hay et al. controlled for the clusters' surrounding environment, so that the only difference between the words /krɛmpɪk/ and /krɛmkɪk/, for example, was the medial cluster. They then compared the ratings and transcription results to the type frequency of each cluster in English. Hay et al. (2003) found that words with high-type-frequency clusters (e.g., /krɛmpɪk/) were rated as more like English words, and low-type-frequency clusters (e.g., /krɛmkɪk/) were not only rated lower, but were often transcribed (incorrectly) as higher frequency clusters. They conclude that their results support models in which word-type frequency is an important determiner of the representation of a phonotactic sequence.

Pierrehumbert (2003b) summarized much of the type-based learning literature to date and used it as a key component of her model of phonological acquisition. In this model,

[6] The extremes of the scale varied across participants so that 7 meant "most wordlike" for half of the participants and "least wordlike" for the other half. Bailey and Hahn used both orthographic and auditory presentations to collect wordlikeness judgments. The results for these different media were essentially the same, and the discussion above refers to both sets of results

discussed in greater detail in Section 2.4.1 below, she combined the storage power of exemplar models and the symbolic systems of generative phonology. The key property of this model is that phonology is a set of learned symbolic representations generalized from word-forms in the lexicon.This model inspired the Edwards, et al. (2004) study of phonotactics and lexical development in which a positive correlation between production accuracy and lexicon size was found. Pierrehumbert's model is also the second of three models used to make predictions for perceptual learning experiments discussed in subsequent chapters.

Type frequency is also part of the inductive, symbolic-rule learning model of Albright and Hayes (2003). Like Bybee (1995, 2001) and Pierrehumbert (2003b), Albright and Hayes contend that type frequency is a necessary component for learning phonology, so models that do not account for type frequency, such as the dual processing model (Pinker, 1999; Pinker & Prince, 1994; Prasada & Pinker, 1993), will not perform as well as models that do. They propose the Minimal Generalization Learner, a model which computes multiple generalizations over word types, as many as can be made from the lexicon. The model sometimes makes generalizations over groups of words that can be classified under a larger generalization, but the more narrow generalization has the advantage of being more predictive for this group of words than it is for the entire lexicon. In support of the claim that people learn both broad and narrow generalizations, Albright and Hayes offer evidence that native English speakers are sensitive to the subclass of regularly inflected verbs that end in voiceless fricatives (cf. "islands or reliability" in work such as Albright, 2002 on Italian verbal morphology). If regular verbal morphology was not sensitive to type frequency, it is difficult to explain why people are sensitive to "especially regular" regular verbs. The Minimal Generalization Learner is the third of three models that were used to make predictions for the dissertation experiments. In the next section I will discuss how Albright (2007) applies the model to phonotactic learning and how the model can make a set of predictions for a set of perceptual learning experiments.

To summarize, type-based learning has been implicated in a large number of phonological studies, including studies of morphological correspondences (Albright & Hayes, 2003; Bybee, 1995, 2001; Pierrehumbert, 2003b) and of phonotactics (Albright, 2007;

Bailey & Hahn, 2001; Pierrehumbert, 2003b). Claims differ from author to author, but the general consensus is that many phonological patterns are abstractions of commonalities between words. This places word learning at the forefront of phonological acquisition generally, and phonotactic acquisition specifically.

I now return to the discussion of how perceptual learning may generalize to production in L1 acquisition. The literature discussed above suggests that *type frequency* is a necessary component of phonological learning. With respect to phonotactics, if a sound sequence is present in many words, then a learner should be able to extrapolate the pattern to new words. If it is present in a few or just one word, however, it is unlikely to generalize to new words. This prediction can be combined with the discussion of perceptual learning. Based on the findings in the L2 literature, type-based models predict that a phonotactic sequence will generalize to production if it appears in multiple word forms in a perceptual input. If a learner hears multiple words containing a phonotactic sequence, such as the clusters /st/ or /fp/ (Gerken et al., 2006), they will learn some representation of that sequence that will facilitate production. If multiple word-types are not encountered, then the pattern cannot be generalized, and no change in productions will occur.

As we will see, however, this is not the only possible means by which phonotactics may be learned. Exemplar-based models, of which Bailey and Hahn's (2001) Generalized Neighborhood Model is just one, do not treat type frequency as fundamentally important. They achieve phonotactic effects by storing individual exemplars, or tokens, of experience, rather than learning explicit representations of phonotactics. Interestingly, the results from the 10 Talkers study provide some support for the exemplar models. This is because children's speed and accuracy at producing a word were influence solely by a manipulation at the token level, that is, talker variability.

The literature reviewed to this point therefore leads us to two conflicting positions. On the one hand, the type-based learning literature suggests that the word level is most critical for learning phonological structures such as phonotactics. On the other hand, exemplar-based models such as Bailey and Hahn (2001) and Goldinger (1998), as well as the results of the 10 Talkers experiments, suggest that it is individual word tokens, or token variability, that is most relevant to learning phonotactics. In the next section I briefly describe a set of experiments in which type frequency and token variability are independently manipulated to determine whether and if each factor contributes to children's ability to learn phonotactics. These experiments can be used to help us better understand whether and how type frequency and token variability contribute to phonotactic learning.

2.4 Experiments, Models of Phonotactic Learning, and Predictions

This section lays out the skeleton of a set of four experiments that test the predictions described above. Following in the methodological tradition of artificial language learning experiments (Gómez & Gerken, 2000) and using the specific methodology described in the Gerken et al. (2006) 10 Talkers study, these experiments simulate perceptually-based language learning in an initial familiarization phase and then test for learning using a production-based task. Children are first exposed to a series of words containing one or more examples of a medial consonant cluster. What differs compared to the 10 Talkers study is that the experiments described below require that children generalize whatever knowledge they have obtained in the perceptual familiarization to the production of a new word with the same medial cluster. In every experiment, at issue is whether the manipulation of Experimental Frequency, or the frequency that a word or cluster appears in the familiarization phase, will influence subsequent production of a related phonotactic sequence. The Experimental Frequency manipulations vary in each experiment, allowing us to determine which manipulations are most relevant to generalization. In Experiment 1, token variability is manipulated and participants were familiarized with clusters spoken by either 1 or 10 talkers (Gerken et al., 2006; Goldinger, 1998). In Experiment 2A, word-type frequency is manipulated, but all words are presented with some level of token variation. In the familiarization phase, each cluster is associated with either one or three word-types, and each word is associated with four talker-tokens. Experiment 2B uses the same methodology as Experiment 2A but a different word set. Finally, Experiment 3 manipulates word-types (one or three words per cluster) without token variation, that is, the same acoustic talker-token is used to present a given familiarization word.

Table 2.2: A short description of the four experiments conducted and presented in this dissertation.

Experiment 1	Experiment 2A	Experiment 2B	Experiment 3
Token variability alone is manipulated	Token variability and type frequency are manipulated	Token variability and type frequency are manipulated	Type Frequency alone is manipulated

I now review in greater depth three of the models described above: the exemplar model proposed by Bailey and Hahn (2001), Pierrehumbert's (2003b) model, and Albright's (2007) Minimal Generalization Learner. The models differ with respect to how type frequency and token variability are treated, and therefore make different predictions about what factors are necessary for generalization. Each model makes a unique set of predictions with respect to the four experiments described above, so the results from the experiments will allow us to choose a model that optimally describes children's performance.

2.4.1 Three Inductive Models of Phonotactic Learning

Exemplar Models

The first class of models that I will discuss, and the class in which token frequency and token variability play the most prominent role, is the exemplar model class (Goldinger, 1998; Hintzman, 1986; Johnson, 1997; Nosofsky, 1986, 1988). Exemplar models have been proposed for a wide range of phenomena, including phonotactics. As such, a substantial literature has been built up about what these models are capable of explaining. Before turning to the Bailey and Hahn (2001) exemplar model of phonotactic effects, I review some of the key proposals in the literature that bear on the architecture proposed by Bailey and Hahn.

Nosofsky's (1986) model was one of the first computationally explicit, exemplar-based proposals of learning and categorization. In this model, exemplars are individual memory traces of perceptual experience stored in memory. An exemplar-based category has no abstract representation, but is determined solely by connections made between the exemplars that constitute the category. Hence, category membership is a function of similarity relationships between the individual category exemplars, rather than their relationship to a category prototype or some other symbolic representation. In this model, "representations" are an epiphenomenal result of activation of related exemplars along relevent dimensions. They do not exist in and of themselves.

Goldinger (1998) made an early attempt at using the exemplar model structure to account for linguistic phoenomena. He used a modified version of Hintzman's (1986) MINERVA 2 model to account for people's memory for talker-specific information. In an earlier experimental study, Goldinger (1996) had shown that people remember a word in a word-identification task better if they had previously been exposed to the word spoken by the same talker. This improved memory for talker-specific tokens was found even when the exposure had occurred one week earlier. To account for this, Goldinger (1998) proposed that linguistic experience is stored as highly detailed exemplars. These exemplars are activated in concert by an incoming word with shared phonetic properties, resulting in a priming effect. Goldinger further suggests that this activation of stored exemplars is what we think of as phonological knowledge. Similarly, Johnson (1997) proposes that speaker normalization is possible because the combined activation of an internally complex category containing information about many speakers is highly flexible and effectively normalizes the incoming signal without a specific process devoted to normalization.

With respect to phonotactics, McClelland and Elman (1986) have proposed the TRACE neural network model of speech perception, which creates associations between words at the phonemic level and can account for phonotactic effects without specifically encoding phonotactic sequences. Bailey and Hahn (2001) follow up on this proposal with an exemplar model that creates phonotactic probability effects via the activation of words that share phonotactic patterns. This model, the Generalized Neighborhood Model, calculates a similarity weight between an incoming word and stored words with shared features. Consider the nonsense word /trɛsp/, an item from Bailey & Hahn's wordlikeness experiments. When a participant in their experiment hears /trɛsp/, words in the lexicon with shared features will activate. The onset /tr/ of /trɛsp/ might create an activation pattern of previously stored words similar to that in Figure 2.3.

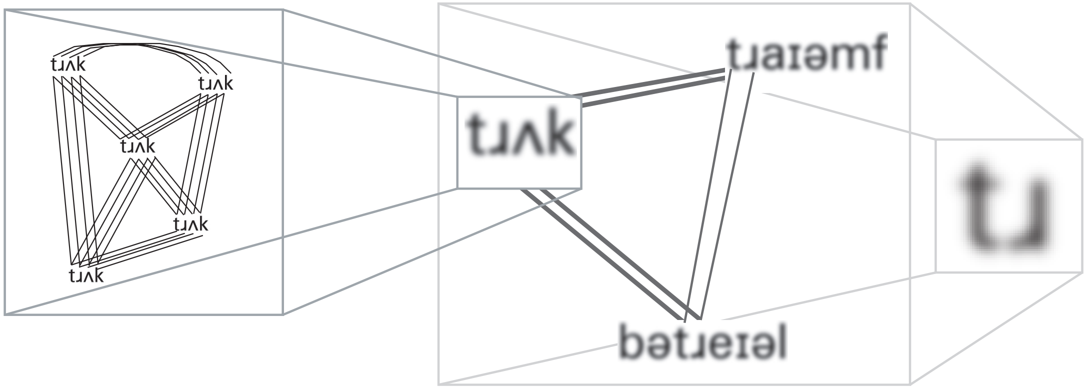


Figure 2.3: An abbreviated model of how activation of multiple exemplars of a word can create the 10 Talker effect and how exemplars of multiple words containing a common sequence, /tr/, can result in a phonotactic effect. Word frequency, or the number of tokens that populate a word category, is encoded in the text and line weights, which represent the connections between exemplars in the inset graphic of *truck* exemplars. *Truck*, which is more frequent relative to *triumph* and *betrayal*, is a category consisting of a greater number of exemplars that form a larger number of connections to related words than do *triumph* and *betrayal*.

The basic level of activation is shown on the left side of the figure, where individual exemplars are linked with one another. These individual exemplars are also connected to exemplars of related words, as shown in the middle of the figure. *Truck*, *triumph*, and *betrayal* are activated because they all share the /tr/ onset in trɛʃp. Note, however, that exemplar models create links between exemplars, not between word types, so the links between words are an idealization of the actual functioning of the model[7]. The square at the far right represents the simultaneous activation of words with a common phonetic sequence, and this activation is what constitutes a phonotactic effect in the model. Phono-

[7] In practice, exemplar models such as Bailey and Hahn's Generalized Neighborhood Model do, in fact, link word types. This implies a certain degree of abstraction which is not present in the Goldinger (1998) and Johnson (1997) models and is somewhat counter-intuitive to the principle of 'categorical behavior by associated exemplars' proposed by Nosofsky (1986). In this description, I assume that the Bailey and Hahn model simplifies the phonetic status of word exemplars, but that the model was intended to represent the type of dynamics described in Figure 2.3.

tactic sequences do not exist in their own right, but are derived entirely from the activation of stored word exemplars that share a common sequence.

As discussed in Section 2.3, Bailey and Hahn's Generalized Neighborhood Model is unable to fully account for adult wordlikeness data. A model of phonotactic probabilities makes an independent and significant contribution to the multiple regression that Bailey and Hahn performed, suggesting that exemplar-based models are insufficient for explaining phonotactic knowledge and that explicit knowledge of phonotactics may exist in the minds of speakers. The exemplar model is nevertheless helpful in creating predictions for the types of phonotactic generalization studies proposed above. Furthermore, exemplar models are ideal for accounting for many experimentally derived frequency effects. Because the models store all, or virtually all[8], experienced input, frequency effects created in an experimental setting are also expected to shape the experimental outcome. For example, the results from the 10 Talkers study fall out naturally from the proposals made by Goldinger (1998) and Johnson (1997).

The left half of Figure 2.3 shows how the 10 Talkers effect can be created in an exemplar model. Incoming word exemplars are stored and compared to each other on a variety of phonetic parameters (e.g., formant values, voice onset time, centroid frequency, etc. See Stevens, 1998 for a detailed description of how phonetic parameters can be encoded in a computational model). As exemplars compile over time, phonetic similarities are encoded through the connections between common phonetic properties across words. By encoding the additional indexical information that identifies a particular talker, the model ensures that each talker's exemplar is stored separately, necessitating the connections between them and allowing for a robust level of activation. Assuming some method of connecting these exemplars to production, such as determining production targets by averaging stored exemplars (Pierrehumbert, 2001), an exemplar model should have no trouble deriving the results from the second experiment in the 10 Talkers study, in which children produced items they heard spoken by 10 talkers (ExpHi) more accurately than the produced items heard spoken by a single talker (ExpLo, cf. Figure 2.2).

[8] See Goldinger, 2007 for a more detailed discussion of storage, particularly as it has been informed by the memory consolidation literature.

I return now to the four perceptual learning experiments described above. In Experiment 1, children are exposed to multiple talker-tokens of one word and are asked to produce a related word. An exemplar model which is highly tuned to token frequency (Skousen, 1989; Nosofsky, 1988), to the phonetic details that distinguish talkers and relate word tokens (Goldinger, 1998; Johnson, 1997), and is accountable for the results of the 10 Talkers study, will predict that generalization should be possible from token variability alone.

To understand why token variability might lead to perceptual learning of a phonotactic sequence, consider an exemplar model that codes incoming tokens of a word at a featural or phonemic level. When the model identifies a test word that shares a phonotactic sequence with the familiarization words, a relatively high similarity score should be possible between the many talker-tokens from the familiarization item, which are linked to one another along various sublexical dimensions, and the test item, which shares a phonotactic sequence. This is because the comparison can include various sublexical dimensions, including phonotactic sequences. This prediction can be altered if, for example, talker-tokens do not link to one another at a sublexical level, although this proposal is not in the general spirit of exemplar models (Goldinger, 1998; Johnson, 1997), and it undercuts the explanation for the results of the 10 Talkers experiment. Therefore, it is safe to say that an exemplar model should predict generalization based on talker-tokens, which distinguishes this type of model from the other two.

In Experiments 2A and 2B, token variability is combined with type frequency. Again, the exemplar model predicts generalization: experience with multiple words containing a phonotactic pattern should lead to improved productions of that pattern in a subsequent production task (Bailey & Hahn, 2001). Talker tokens of a given word will be linked by the phonetic characteristics of the word, and the different words will be linked by their common phonotactic sequence. This sequence should be activated again during the production test, leading to improved production.

With respect to Experiment 3, exemplar models predict generalization effect here, as well. Links across word-types at the sublexical level should result in the same activation of the target phonotactic sequences expected in Experiments 1, 2A, and 2B (Bailey &

Hahn, 2001). This means that the exemplar model predicts a generalization result in all four experiments. In every case, some form of variation (talker variation in Experiments 1, 2A, and 2B; word-type variation in Experiments 2A, 2B, and 3) will allow for connections between the various exemplars sharing a phonotactic sequence, and the related test sequence will activate those sundry connections. This set of predictions is unique to exemplar models, as the other models explicitly require multiple word-types for generalization to occur.

One problem with this set of predictions is that it is cobbled together based on the properties of several exemplar models. The predictions are meant to represent an ideal model, that is, one which stores and links highly detailed exemplars. Should the predictions, fail, however, no single model can be held accountable. This problem, rather than reflecting a weakness in the experimental design, represents a problem with exemplar models as a class. No rigorous definition of *exemplar* exists, nor does *detail* as it relates to the internal composition of exemplars (cf. Bailey & Hahn, 2001 for multiple proposals about what forms the internal composition of exemplars could take). These criticisms have been raised elsewhere (cf. Goldinger, 2007), but they are worth noting here because it is not possible to manipulate token variability or type frequency such that the model predicts "no generalization." In other words, the model can only be challenged by arguing from a null effect, calling into question the falsifiability of the entire class of models. Given this concern, it seems reasonable to test the model as defined. This definition conforms to general principles of exemplar storage and, as we will see, contrasts with the predictions made by the other two models. With this set of predictions established, I turn now to Pierrehumbert's (2003a, 2003b) model.

Pierrehumbert's Hybrid Model

The exemplar model's general properties are not necessarily in line with other facts about language. The difficulties that pure exemplar models have in explaining certain linguistic facts is discussed at length by Pierrehumbert (2006, see also Nosofsky and Bergert (2007), Nosofsky and Zaki (1998), Smith (2005), and Smith and Minda (2002) for non-linguistic criticism). Most problematic is that levels of abstraction have been implicated in studies of the different processing effects of neighborhood density and phonotactic probability (Vitevitch & Luce, 1998; Vitevitch, Luce, Pisoni, & Auer, 1999; Bailey & Hahn, 2001). The results from Bailey and Hahn's (2001) and Albright's (2007) wordlikeness experiments, for example, suggest that both neighborhood density *and* phonotactic probability make independent contributions to people's judgments of wordlikeness. For this reason, Pierrehumbert (2003a, 2003b) suggests that exemplar models are valuable for explaining certain low level processing effects, but that abstraction is necessary to explain higher order phonological knowledge, such as knowledge of phonotactics. This is the hybrid component to her model. Children do not simply store unanalyzed chunks of language. Patterns that arise through statistical modes in parametric phonetic space are instantiated as abstract linguistic representations. These abstractions are entirely dependent on linguistic experience, however, and are language specific.

In many ways, Pierrehumbert's model was designed specifically to account for the phonotactic frequency facts discussed earlier (cf. Section 2.1.2). Much of the literature covered here is also covered in Pierrehumbert (2001, 2003a, 2003b), particularly frequency effects in perception and production. The correlation between lexicon size and production accuracy found by Edwards et al. (2004) was based on the prediction made by Pierrehumbert that such a correlation should exist. Type frequency, too, is assumed to have an effect in this model.

Only perceptual learning is left out. Although Pierrehumbert does not attribute the frequency effects seen in child speech production directly to perceptual learning, it does appear to be implied. Category learning, for example, is based on the perceptual learning mechanism described in Maye and Gerken (2000) and Maye, Werker, and Gerken (2002), and no articulatory feedback mechanism is mentioned or implied. Therefore, it is safe to say that generalization of phonotactics from a perceptual learning mechanism is line with the general theoretical outline of Pierrehumbert's model.

The predictions made by Pierrehumbert's model can be summarized as follows: word-type frequency is required for generalization of a phonological pattern, such as a phonotactic sequence, so a null effect of Experimental Frequency is expected for Experiment 1, making the model distinct from the exemplar model described above. Experiments 2A and 2B, which do exploit word-type frequency, are predicted to reveal significant effects of Experimental Frequency. Pierrehumbert's model uniquely predicts a null effect for Experiment 3, which removes token variability. The reasons for this prediction are explained now.

A critical difference between the hybrid model and other exemplar models is that abstract representations do exist. These representations are built bottom-up from raw perceptual experience stored in memory. From there, generalizations are made based on reliable statistical patterns, which are converted to abstract, symbolic units. Crucially, abstractions may be built upon abstractions. Phonotactics, for example, are abstractions defined over lexical items, which are themselves abstractions. This prediction can be seen in the following quote from Pierrehumbert (2003b).

> Each word in a speaker's lexicon has a representation of its sound structure that allows it to be recognized despite variability in its phonetic form resulting from speaker differences and context. The same representation presumably mediates between perception and production, making it possible for speakers to repeat words they have acquired through perception. Given this description, it is clear that word-forms are also abstractions over phonetic space. (p. 179)

Considering the data from Gerken et al. (2006), Pierrehumbert's interpretation of the results is clear: variability at the parametric phonetic level is necessary for an abstract lexical entry to be stored. Once it has been stored, it can mediate between perception and production, leading to the increased speed and accuracy that were seen in children's productions of words spoken by 10 talkers. Figure 2.4 shows how exemplars stored from multiple talkers leads to an abstract representation of a word. The abstract representation of the word /mæfpəm/ is linked to both the perceptual and articulatory systems, resulting in the facilitative effects seen in the 10 Talkers study.

Figure 2.5 shows how representations built from the bottom up can lead to abstract knowledge of a phonotactic sequence. Multiple talker-tokens lead to abstract word forms, which themselves can be used to create a generalization. In this case, all of the words share a medial cluster, /sp/, which lead to a corresponding generalization. From this description, we can see that Pierrehumbert's model predicts a null effect in Experiment 3, in which variability at the lowest level (talker variability noted as T1, T2, etc. in Figure

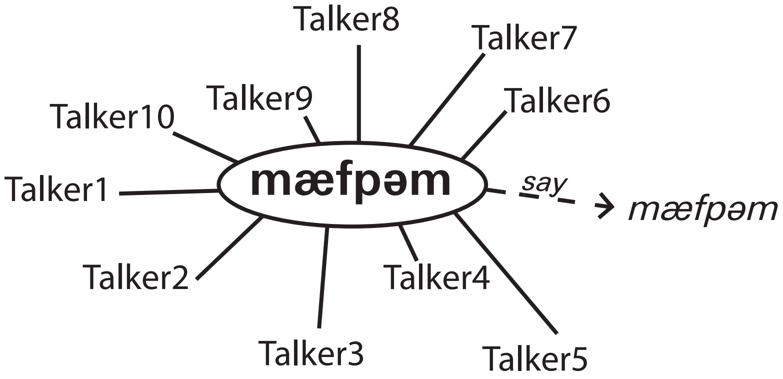


Figure 2.4: Creation of an abstract word form according to Pierrehumbert's (2003a, 2003b) hybrid model. Phonetic level variability, such as different talkers, leads to abstraction of word forms.

2.5) is absent. Without the phonetic variability derived from multiple talkers, children in that experiment should be unable to create abstract lexical items and should therefore be unable to learn anything about phonotactics.

A weakness of the hybrid model is that, compared to the models described by Bailey and Hahn (2001) and Albright (Albright, 2007; Albright & Hayes, 2002, 2003), it is the least mathematically explicit. The other models provide equations that can be used to fit predictions made by the model to human data. Most problematic, the transition from the low-level exemplar model to an abstract phonological grammar, or the transition from an unsupervised to a supervised learning algorithm, is a very challenging computational task. The prose description of the model does offer several predictions with respect to the perceptual learning experiments, however, and this set of predictions is unique to Pierrehumbert's model. I will now describe Albright's MGL and use its architecture to make predictions for the four perceptual learning experiments.

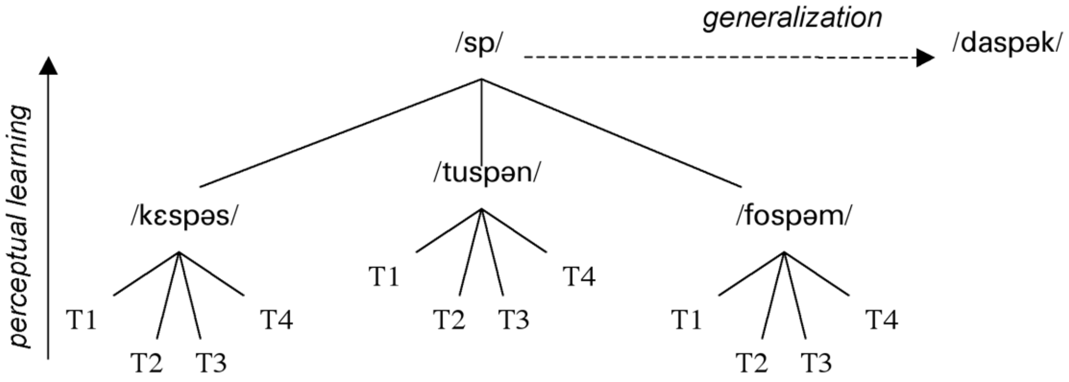


Figure 2.5: Creation of an abstract representation of the phonotactic sequence /sp/ according to Pierrehumbert's hybrid model. Following the creation of abstract word forms from low-level phonetic variability, word-type variability leads to abstraction of phonotactics and other phonological structure.

Albright's Minimal Generalization Learner

The MGL (Albright, 2002; Albright & Hayes, 2002, 2003; Albright, 2007) was originally created to account for judgment data related to morphological correspondences. For example, Albright and Hayes (2003) asked people to create a past tense form (Berko, 1958) for the word *sliff* and then rate a number of possible forms (Prasada & Pinker, 1993) for *sliff*, including *sliffed* and *sluff* . Generally, people use the regular past tense suffix *-ed* for most nonsense verbs, but use an irregular past tense formation process, such as *sliff-sluff*, when the verb is phonologically similar to other verbs that alternate in the same way.

What makes the MGL unique compared to other symbolic-processing models is that it learns regular and irregular past tense formation processes by the same mechanism. The model works from word pair to word pair, iteratively making more and more general rules from the initial base pairs. The model looks for a *structural change* to describe how one member of the pair is derived from the other, as well as a *context* in which the change takes place. Essentially, these are rules of the type formalized by Chomsky and Halle (1968)[9]. The model compares structural changes, and when it finds a match, it attempts to create a more general rule (by dropping out phonological features from each

[9] Albright and Hayes (2003) suggest that a more nuanced formalism could be created based on research over the last 40 years, but leave the implementation to future research.

rule's respective context). An important difference between this approach and that of Chomsky and Halle and other generative models is that the Albright and Hayes model seeks minimal generalizations and allows more specific rules to be redundant with more general rules that create the same alternation. In other words, it stores the less general rules and often ends up with multiple rules that describe the same structural change but have different contexts. Albright and Hayes (2003) find that the MGL performs quite well when fitting people's past tense formations, as well as their intuitions about which formation strategies are best, given the choice between several.

In a recent manuscript, Albright (2007) extends the MGL to acceptability judgments of phonotactics. In congruence with the research on frequency effects above (Section 2.1), Albright notes that people's judgments of nonsense words are closely tied to the statistics of phonologically related real words. To explain this fact, he advocates a combinatorial grammar which, "makes use of knowledge of combinatorial phonotactic possibilities of different sounds in the language" (p. 4). A visual example of how the model works can be seen in Figure 2.6. It makes minimal generalizations over phoneme pairs and extracts shared features, leading to generalizations about natural classes. He contrasts this with neighborhood activation models (Luce, 1986; Bailey & Hahn, 2001), and models that compute only phonotactics, such as Vitevitch & Luce's (2004) Phonotactic Probability Calculator (which calculates summed biphone phonotactic probabilities). He further contrasts his model with joint transitional phonotactic probability models (which calculate the products of a word's component phonotactic probabilities).

In fitting each model to data from Bailey and Hahn (2001) and Albright and Hayes (2003), Albright finds that Bailey and Hahn's model works best for their own data, and a joint transitional biphone probability model and the MGL perform equally well for the data from Albright and Hayes. Albright notes several advantages to the Albright and Hayes data, however, including a wider range of phonotactic probabilities and a procedure for removing perceptual errors from judgments. Following Bailey and Hahn (2001), he concludes that there is strong support for a grammatical component to acceptability judgments, but reinterprets the grammatical component as being the most relevant.

Like Pierrehumbert's model, Albright's MGL is designed to account for phonotactic

	#	t	ɹ	ʌ	k	#
		$\begin{bmatrix}+\textbf{ cons}\\ -\textbf{cont}\\ +\textbf{cor}\\ -\textbf{voi}\end{bmatrix}$	$\begin{bmatrix}+\textbf{cons}\\ +\textbf{approx}\\ +\textbf{cor}\\ +\textbf{voi}\end{bmatrix}$	$\begin{bmatrix}+\textbf{syll}\\ +bk\\ \ldots\end{bmatrix}$	$\begin{bmatrix}+cons\\ \ldots\end{bmatrix}$	
b	ə	t	ɹ	eɪ	ə	l
	$\begin{bmatrix}+syll\\ +bk\\ \ldots\end{bmatrix}$	$\begin{bmatrix}+\textbf{cons}\\ -\textbf{cont}\\ +\textbf{cor}\\ -\textbf{voi}\end{bmatrix}$	$\begin{bmatrix}+\textbf{cons}\\ +\textbf{approx}\\ +\textbf{cor}\\ +\textbf{voi}\end{bmatrix}$	$\begin{bmatrix}+\textbf{syll}\\ -bk\\ \ldots\end{bmatrix}$	$\begin{bmatrix}+syll\\ +bk\\ \ldots\end{bmatrix}$	$\begin{bmatrix}+lat\\ \ldots\end{bmatrix}$

Figure 2.6: An example of the distinctive feature search made by the MGL to learn the phonotactic sequence tr[+syll] from the words *truck* and *betrayal*. The words are compared for common feature sets, and the minimally general, that is, the most specific generalization true of both words is made. Features that are part of the generalization are given in bold print. The symbol # represents a word boundary.

probability effects. It generates a probabilistic grammar that adapts to the statistical properties of the lexicon, so it is well suited to handle phonotactic frequency effects in adult wordlikeness judgments (Albright & Hayes, 2003; Bailey & Hahn, 2001). It has not been used to model phonotactic probability effects seen in child production data, however. To date, the MGL has only been tested on adult judgment data. We might therefore question the model's ability to handle facts such as the the correlation between vocabulary size and production accuracy found by Edwards et al. (2004). The model also lacks a perception-production link. These concerns are not problems in principle, however, and the model could be adapted to account for these facts (Albright, 2008, personal communication). This adaptation requires only one assumption: that speech production tasks tap phonotactic knowledge. To the extent that phonological representations inform the functioning of the speech production system, and to the extent that the production mechanism connecting the grammar and the motor system is transparent, the MGL will predict that perceptual learning can lead to generalization.

The MGL is quite clear about requiring multiple word-types for making generalizations, so it predicts a null result for Experiment 1, distinguishing the MGL from the exemplar models. What separates the MGL's set of predictions from Pierrehumbert's model is that it does not make a distinction between Experiments 2A, 2B, and 3. The MGL is preequipped with phonological features, and does not require learning at the phonetic level or learning of abstract lexical entries. Lexical entries are stored as distinctive features/phonemes. Therefore, while it predicts no learning is possible in Experiment 1, it also predicts no difference in learning between Experiments 2A, 2B, and 3. That is, the MGL predicts Experimental Frequency to show effects in the final three experiments.

The predictions for the three models across all four experiments are recapitulated in Table 2.3. Although all three models make the same predictions for some experiments (2A and 2B), the combination of all four experiments provides a set data which allow the three models to be compared. The model whose predictions most closely match the results of these experiments therefore represents the optimal model of phonotactic learning.

Table 2.3: Predictions made by the three models—Bailey and Hahn's exemplar model, Pierrehumbert's hybrid model, and Albright's MGL—for the experiments conducted for the dissertation. *Gen* is given in cells where a model (row) predicts a significant effect of the Experimental Frequency manipulation for a given experiment (column). *No Gen* refers to a predicted null effect. A prediction that is unique to a particular model is boxed and in bold.

	EXP 1	EXP 2A	EXP 2B	EXP 3
	Tokens	Types + Tokens	Types + Tokens	Types
EXEMPLAR MODEL Bailey and Hahn (2001)	**Gen**	Gen	Gen	Gen
HYBRID MODEL Pierrehumbert (2003b)	No Gen	Gen	Gen	**No Gen**
MINIMAL GENERALIZATION LEARNER Albright (2007)	No Gen	Gen	Gen	Gen

Chapters 3, 4, 5, and 6 present the results of Experiments 1, 2A, 2B, and 3, respectively. The results of each experiment will be described as they relate to the predictions of the three models. Chapter 7 provides a general discussion of the results and offers ideas for future research that will provide additional tests of the models, as well as practical applications of the perceptual learning procedure. Chapter 8 concludes the dissertation with a review of the findings and a general discussion.

CHAPTER 3

EXPERIMENT 1: PHONOTACTIC GENERALIZATION FROM MULTIPLE TALKER-TOKENS

As mentioned in Chapter 2, Gerken et al. (2006) found that four-year-old children produce nonsense words faster and more accurately when they were first familiarized with 10 tokens of the word spoken by different talkers. This finding was true for words containing both high and low probability phonotactic sequences. However, that experiment did not test whether children can *generalize* the target phonotactic sequences to new words. To test whether perceptual familiarization effects generalize, the experiments reported in this dissertation consider whether perceptual familiarization with a phonotactic sequence in one set of words can generalize to productions of the same sequence in a set of new words.

The focus of this chapter is Experiment 1, in which four-year-old children were tested on their ability to generalize phonotactic information based on token variability. The children were first exposed to the target phonotactic sequences during a perceptual familiarization. The phonotactic sequences were eight consonant clusters, four High English Frequency and four Low English Frequency, embedded medially in CVCCVC nonsense words. In the subsequent production test, children were asked to repeat the same clusters in words that they had not heard previously. The results were analyzed for accuracy and speed (production latency).

Existing literature shows that children produce high probability phonotactic sequences more accurately than low probability sequences (Beckman & Edwards, 1999; Edwards et al., 2004; Munson, 2001; Zamuner et al., 2004), which led to the prediction that the HiEng clusters would be produced more accurately than the LoEng clusters.

The familiarization phase also included a manipulation of Experimental Frequency: half of the familiarization words were played only once, the others played 10 times, with a different talker producing each token (Gerken et al., 2006). As discussed in the previous chapter, different models of phonotactic learning make different predictions about the Experimental Frequency manipulation, or the frequency with which a cluster was heard in the experiment. These predictions, as laid out in Chapter 2, are repeated in Table 3.1 below. Exemplar-based models are unique in predicting that generalization is possible based on token variability alone. This is because exemplar models are uniquely sensitive to phonetic details such as talker variability (Goldinger, 1998; Johnson, 1997) and are ideal for accounting for data such as that found in the 10 Talkers experiments. The sublexical associations that allow exemplar models to explain the 10 Talker effect should also allow the model to make associations to a related word, resulting in a generalization effect. The models proposed by Albright (2007) and Pierrehumbert (2003b) are instead reliant on abstract lexical items, or type frequency, to learn phonotactics, and they predict a null effect of Experimental Frequency: participants should not be influenced by the token frequency of the familiarization words when they produce the related test words.

Table 3.1: Predictions made by the three models—Bailey and Hahn's exemplar model, Pierrehumbert's hybrid model, and Albright's MGL—for the token variability experiment. *Gen* is given in cells where a model (row) predicts a significant effect of the Experimental Frequency manipulation for a given experiment (column). *No Gen* refers to a predicted null effect. A prediction that is unique to a particular model is boxed and in bold.

	EXP 1 Tokens	EXP 2A Types + Tokens	EXP 2B Types + Tokens	EXP 3 Types
EXEMPLAR MODEL Bailey and Hahn (2001)	**Gen**	Gen	Gen	Gen
HYBRID MODEL Pierrehumbert (2003b)	No Gen	Gen	Gen	**No Gen**
MINIMAL GENERALIZATION LEARNER Albright (2007)	No Gen	Gen	Gen	Gen

3.1 Method

3.1.1 Participants

Twenty-five children between 48 (4;0) and 52 (4;3) months of age were recruited for the study (mean = 4;1.27). Children were recruited using a database of local birth announcements and adoptions in the Tucson metropolitan area. As a control for linguistic development, the parents/caretakers filled out a questionnaire regarding the child's language history. With respect to second language exposure, all parents reported that their children were native English speakers. When children had been exposed to a second language, parents were asked to report the number of hours of exposure per week. Any children whose weekly exposure was greater than 10 hours were excluded from the analysis. All of the participants had minimal exposure to another language, however. Across all participants, the average exposure to a second language was 1.43 hours. Two children were exposed to Spanish for eight hours per week, one child was exposed to Hebrew for two hours per week, and one child was exposed to Chinese for five hours per week. All other participants were reported to have no regular second language exposure. With respect to pathological language development, 23 children were reported to have had no personal or family history of any of the following: early intervention services, ear infections in the month prior to their participation, congenital hearing loss, congenital language delay, speech or language therapy. Two children whose parents reported atypical language development were removed from the analysis. Six other children were removed from the analysis because they did not complete the experiment, and one child was removed because of an experimenter error. The remaining 16 participants, 8 girls and 8 boys, were included in the analysis discussed below.

3.1.2 Materials

Consistent with the studies conducted by Gerken et al. (2006) and Munson (2001), all words were CVCCVC nonsense words, and are given in Table 3.2. The words were created in pairs, with eight pairs total. The members of a pair shared a common phonotactic sequence, their medial cluster. One member of the pair was assigned to the familiarization word set, the other member was assigned to the test word set. The familiarization words were used to introduce the phonotactic sequences and were the same words used in the 10 Talkers study (cf. Table 2.1). The test words were used during the production task and were created specifically for this experiment. Novel words were used so that the effects of the familiarization could be seen without the influence of factors such as word frequency and imageability.

Four word pairs contained High English frequency clusters (HiEng); the other four pairs contained infrequent or unattested clusters of English (LoEng). This division is the English Frequency factor. The HiEng and LoEng words were matched such that the same word frame was used for both a HiEng and a LoEng word. For example, the frame /mæ__əm/ was combined with the clusters /st/ and /fp/ to create the words /mæstəm/ and /mæfpəm/.

Table 3.2: Eight familiarization words and eight test words used in Experiment 1. The familiarization and test sets are in different columns. Matched HiEng and LoEng words that share a frame are given in corresponding rows for their respective clusters. For example, the familiarization items /fospəm/ and /foʃpəm/ share a word frame and are both listed in the first row of each set of clusters. The matched test words /daktən/ and /dapkən/ are given in the fourth row. Words are written in IPA.

	Familiarization Words	**Test Words**
HiEng CC	fospəm	zaspən
	mæstəm	neɪstən
	fæmpɪm	simpən
	boktəm	daktən
LoEng CC	foʃpəm	zaʃpən
	mæfpəm	neɪstən
	fæmkɪm	simkən
	bopkəm	dapkən

Several factors made consonant clusters the ideal sequences for examination. First, clusters are relatively well studied as a developmental milestone, both in normally developing (Locke, 1983; McLeod et al., 2001; Olmsted, 1971; Prather et al., 1975; Smit,

1993) and impaired (Chin & Dinnsen, 1992; Elbert & Gierut, 1986; Hodson & Paden, 1981; Shriberg & Kwiatowski, 1980) populations. Second, accurate cluster production develops relatively late (McLeod et al., 2001; Smit, 1993; Smit, Hand, Freilinger, Bernthal, & Bird, 1990). Smit (1993) showed, for example, that the frequent clusters /st/ and /sp/ are produced accurately in word-initial position by only 75% of 4 1/2 year-olds. This finding, based on a large survey of children learning English in the United States (Smit et al., 1990), led to the expectation that children's productions in the present experiment should contain a consistent number of inaccuracies.

Another reason for using clusters is that two-consonant clusters of the type used here are biphones and are therefore an ideal unit for the study of phonotactic learning[1]. Finally, using clusters as the target phonotactic sequence allowed for comparisons with the results obtained by Beckman and Edwards (1999), Munson (2001), Edwards et al. (2004), and Gerken et al. (2006). All of these studies used the same set or a similar set of word-medial consonant clusters and varied the phonotactic probability of those clusters. In fact, the clusters used in Gerken et al. (2006) and in the studies presented here are a subset of the clusters used by Munson (2001). In summary, clusters are well-studied phonotactic sequences, and their use facilitates a comparison to previous findings related to phonotactic probability effects in child speech.

The word list was kept small for two reasons. First, it kept the length of the familiarization task under two minutes, a manageable time for young children. Second, it allowed for the collection of four repetitions of each test word from each child without straining their patience. This restricts a by-items analysis because very little power can be achieved with only eight unique data points, which might be conceived of as a limitation of the design. In fact, a power estimate based on the F-value of a by-items analysis of the accuracy results from the 10 Talkers study (where $F = 2.15$) suggests that a by-items analysis for the data from this experiment will have a power, $\phi \approx 0.1$, far smaller than the standard value of 0.8.

The widespread use of by-items analyses notwithstanding, an items based analysis

[1] Pierrehumbert (2003b) showed that the average adult lexicon is not large enough to allow for robust categorization of triphone sequences. She goes on to argue that phones and biphones are likely the only abstract phonotactic sequences represented in the grammar.

is not necessarily appropriate for this type of study. The basic assumption behind a by-items ANOVA is that items are randomly sampled from the larger population of possible lexical items, just as participants are thought to be an accurate sampling of the larger human population. The nonsense words used as items in this experiment are not random, however. Word frames were used to match HiEng and LoEng clusters, creating highly similar words for both English Frequency conditions, and the same words were used for the Experimental Frequency variable, a manipulation of the number of talkers heard producing a familiarization word. The important statistical test here is therefore the by-subjects analysis, and it is the primary test that is discussed below. For further discussion on the relevance of by-items analyses, see Raaijmakers, Schrijnemakers, and Gremmen (1999)[2]. In the next few sections, the English Frequency and Experimental Frequency factors are described in more detail.

English Frequency

As shown in Table 3.2, the clusters were divided according to their frequency in English. The High English Frequency (HiEng) clusters were /kt/, /mp/, /sp/, and /st/; the Low English Frequency (LoEng) clusters were /pk/, /mk/, /ʃp/, and /fp/. The choice of clusters and their partitioning follows a study on the production of consonant clusters by Munson (2001), who used a larger set of clusters that included the clusters used here[3]. The online Phonotactic Probability Calculator (Vitevitch & Luce, 2004, www.people.ku.edu/~mvitevit/PhonoProbHome.html) was used to determine the phonotactic probabilities of the clusters. It is determined by taking the $\log_{10}$ value of the quotient of the total number of words in the language that contain a phonotactic sequence over the total number of words in the lexicon. Phones, or the probability of a single phoneme at a given point in the word, and biphones, the probability of a two-phoneme sequence, are

[2] Another alternative to the by-items analysis is the multilevel model, an analysis allowing for hierarchical or nested factors such as subjects and items, and which is discussed by Baayen (2004) as it relates to psycholinguistic experimentation. Although this is certainly a valid solution, it is not addressed here.

[3] In this study, the clusters were partitioned according to their phonotactic probability (Jusczyk et al., 1994; Vitevitch & Luce, 1998). Munson refers to his measurement as a measure of "transitional probability," although the description of the measurement is virtually identical to the biphone probability measurement described by Vitevitch and Luce and used here.

the usual phonotactic probabilities of interest (cf. footnote 1 of this chapter). The biphone probabilities of the clusters used in this experiment are given in Table 3.3. The Phonotactic Probability Calculator calculates phonotactics with respect to word boundaries, but not syllable structure or stress Therefore, the probabilities in Table 3.3 give the likelihood of that consonant sequence appearing somewhere in the middle of a word. For example, /st/ is the most frequent cluster and has a probability of 0.0232 of occurring word-medially, whereas /ʃp/ and /fp/ never occurred in the corpus and have probabilities of 0.0. To determine whether there was a significant difference between the two groups of clusters, an ANOVA was conducted on their biphone frequencies. Scores for the HiEng clusters were significantly higher than scores for the LoEng clusters (*F* (1,6)=6.611, $p < .05$). The results suggest that partitioning phonotactic probability into two discrete categories was justified[4].

Table 3.3: Biphone probabilities for the eight clusters used in Experiment 1.

	HiEng CC				**LoEng CC**			
Cluster	/kt/	/mp/	/sp/	/st/	/pk/	/mk/	/ʃp/	/fp/
Biphone Probability	0.0036	0.0091	0.0081	0.0232	0.0002	0.0002	0.0	0.0

To ensure that the phonotactics were balanced across the familiarization and test word sets, paired t-tests were conducted on the phone and biphone sums corresponding to each word. A *phone sum* is the sum of all the component phones of a word, a *biphone sum* is the sum of all the component biphones of a word. The phone test was not significant (*t* (7)=-1.801, *p* = .115), nor was the biphone test (*t* (7)=-1.200, *p* = .269). The analysis shows that the two word sets were reasonably well balanced for whole-word phonotactic probabilities. Table 3.4 gives the phone and biphone sums for the familiarization and test words used in Experiment 1.

[4] There is a possibility that additional effects could be found by a finer grained division of English Frequency, as exists between the very high frequency cluster /st/ and the other HiEng clusters. A more detailed exploration of English Frequency, particularly as a continuous factor, remains for future research.

Table 3.4: Sums of phone and biphone probabilities for the familiarization and test words used in Experiment 1.

	Familiarization Words			**Test Words**		
English Freq	*Word*	*Phone Sum*	*Biphone Sum*	*Word*	*Phone Sum*	*Biphone Sum*
High	boktəm	0.3605	0.0259	daktən	0.42	0.0319
	fæmpɪm	0.2918	0.0195	simpən	0.3847	0.0342
	fospəm	0.328	0.0325	zaspən	0.343	0.0317
	mæstəm	0.4219	0.0577	neɪstən	0.3861	0.0505
Low	bopkəm	0.297	0.0171	dapkən	0.3565	0.0238
	fæmkɪm	0.2978	0.0095	simkən	0.3907	0.023
	foʃpəm	0.2569	0.0229	zaʃpən	0.2719	0.0214
	mæfpəm	0.3096	0.0272	neɪfpən	0.2738	0.0232

Syllable structure One concern that might be raised about the English Frequency factor is that it is not really a division of high and low probability sequences, but instead it is a division of licit/illicit syllabic units. A search of the MRC Psycholinguistic database (http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm) yields the following facts: the HiEng clusters /sp/ and /st/ are both licit word onsets (*spoon* and *store*, for example), and all four HiEng clusters are licit codas (*cusp*, *dust*, *dump*, and *duct*). For the LoEng clusters, although, /mk/ and /pk/ do appear in English words such as *tomcat* and *napkin*, respectively, most English speakers would agree that the /k/ of each cluster is the onset of a new syllable. The other LoEng clusters /ʃp/ and /fp/ are extremely rare and have the same syllabic properties as the other LoEng clusters, although /ʃp/ does occur medially in words like *flashpoint* and appears as an onset in a few words borrowed from Yiddish, including *spiel* and its inflected forms. It is nevertheless reasonable to conclude that none of the LoEng clusters are allowable as onsets or codas in English. The issue, then, is whether the English Frequency factor should be interpreted at all with respect to phonotactic frequency, or whether it would be better to interpret it as reflecting syllable structure differences.

An empirically motivated reason to interpret the English Frequency factor as being about phonotactics comes from previous studies of probabilistic phonotactics. These studies, employing regression analyses, have shown significant effects of phonotactic frequency in speech production measures, regardless of syllable structure. Edwards et al. (2004), for example, used a regression analysis to determine the effect of phonotactic frequency on the average accuracy scores for their words with the phonotactic frequency of the target sequence in that word, and found a significant effect: words were produced more accurately when they had higher phonotactic probabilities. As in the present study, Edwards et al. used medial consonant clusters with various syllable structure properties, but nevertheless found that children's productions of those clusters correlated with their phonotactic probabilities.

Similarly, Hay et al. (2003) showed that adults are likely to mistake a low probability nasal-obstruent cluster with a higher probability cluster, and that adults rate words with high probability clusters as more word-like than low probability clusters. Like the results from the Edwards et al. (2004) study, a regression analysis found a significant contribution of phonotactic probability to the measurement of cluster accuracy in spite of differences in syllable structure. Notably, the word set used by Hay et al. included the clusters /mk/ and /mp/, which are being used here. The Edwards et al. (2004) and Hay et al. (2003) studies confirm that phonotactic probabilities are relevant, even when syllable structure may differ.

In fact, it may not be possible to independently manipulate syllable structure and phonotactic probability in a study such as this one. As mentioned above, clusters were chosen at the poles of the phonotactic probability continuum. Although this allowed for the inclusion of phonotactic probability as a discrete, rather than a continuous, factor, it also forced the syllable structure confound. High probability clusters occur frequently in words because they can occur as onsets or codas; low probability clusters occur infrequently or not at all because they do not. It appears, then, that this confound was largely unavoidable.

It should be noted that the phonetic qualities of the materials bias an interpretation where the first consonant is considered a coda and the second an onset. As discussed in the next section, all the stop consonants occurring in the second position, such as the /p/

in /dɪmpət/ and the /k/ in /dɪmkət/ were produced with some aspiration[5]. Phonologically, this is most consistent with onset status, which coerces the status of both the HiEng and LoEng clusters as being split by a syllable boundary. It might be best to think of the HiEng and LoEng clusters as *consonant sequences*, which is neutral to syllable structure, if not slightly less informative than *cluster*. I use the term *cluster* henceforth as a matter of convenience, but also because the results from Gerken et al. (2006) suggest that consonant sequences of different syllabic structure behave similarly with respect to perceptual learning.

The issue of syllabification remains an important part of phonological knowledge, however, so it should not be dismissed as irrelevant. It is reasonable to assume that both phonotactic probability and syllabification factors influence children's productions of a novel word, and that both will influence the results of this and the subsequent experiments. The issue of syllabification will return at various points in the dissertation, and, where relevant, I will attempt to interpret its significance in the results obtained.

As a final note on the English Frequency factor and its interpretation as distinguishing phonotactic probabilities, or syllabic structures, or both, it should be noted that this factor in not of primary interest here. In fact, the English Frequency factor is included largely as a control measure and a way to compare the results to previous studies. The central research question is about how children's production speed and accuracy are affected by frequency effects created *within* the experimental design. This is the question to be addressed by the Experimental Frequency factor, which is discussed next.

Experimental Frequency

In addition to manipulating the frequency of phonotactic sequences as they occurred in English, the frequency that children heard them in the experiment was also manipulated, what I will refer to as Experimental Frequency. During the familiarization phase of the

[5] This quality of the items also reduces the chances that the nasal-stop clusters were interpreted as nasal vowel + stop consonant, as is often the phonetic reality for nasals when they occur syllable-finally (Kenstowicz, 1994). The concern that nasals might not be produced as consonants was also addressed by using a system of wide transcription, so that nasalized vowels were interpreted as word-final nasals. See the Analysis section below for further details.

experiment, children heard Experimental Frequency High words (ExpHi) 10 times each and the Experimental Frequency Low words (ExpLo) only once. Following Gerken et al. (2006), a different talker was recorded for each of the 10 ExpHi word-tokens.

All familiarization words were recorded by 10 women, all native speakers of American English with a Western United States accent[6]. The recording took place in a sound attenuated booth, using an Andrea anti-noise USB NC-7100 microphone, and were digitized using Sound Studio software (www.freeverse.com/soundstudio/) on an Apple iMac. Talkers were asked to produce several tokens of each word carefully and to provide clear cues for the component phones, particularly the two phones forming the medial cluster. For example, talkers were asked to release the first stop of the cluster and aspirate the second stop[7]. The words were produced with stress on the first syllable, or trochaic stress. From these recordings, a single production of each word from each speaker was chosen to function as a familiarization token. The selected productions were extracted from the recording with 100 ms of silence on either side, and were normalized for root-mean-square amplitudes.

The familiarization words were divided into two presentation lists[8], so that an word was ExpHi in one list and ExpLo in the other. In the first list, the words /boktəm/, /fæmpɪm/, /bopkəm/, and /fæmkɪm/ were ExpHi, and /fospəm/, /mæstəm/, /foʃpəm/, /mæfpəm/ were ExpLo. The opposite assignment was used for the other list. These lists

[6] The use of a single gender follows from research showing that infants have greater difficulty recognizing known words spoken by talkers of a different gender (Houston & Jusczyk, 2000) than by talkers of the same gender. Given the facilitative view of talker variability found in Houston and Jusczyk (2003) and adopted here, however, it is possible that using a familiarization item set with talkers of both genders might prove to be more facilitative that an item set comprising talkers of a single gender. This possibility, while worth exploring, is not addressed here.

[7] Phonetically, syllable-final stops are regularly unreleased in English. Stops following fricatives, as in the clusters [st] and [sp] are often unaspirated and therefore indistinguishable from [d] and [b], respectively. Requiring the first stop to be released and the second to be aspirated provided an additional level of balance between the HiEng and LoEng clusters, which differ with respect to syllable structure (cf. previous section). The fact that the first stop in every cluster was released and the second was always aspirated should have reduced syllable structure related differences between the HiEng and LoEng clusters.

[8] The term *set* is used to refer to the Familiarization and Test word groups. The term *list* refers to the division of clusters into ExpHi for one list and ExpLo for another list. For example, the cluster /sp/ was represented by the word /fospəm/ in the *Familiarization set* and by the word /zaspən/ in the *Test set*. Looking at /fospəm/, this word was spoken by just one talker, that is, /fospəm/ was ExpLo, in *List 1*, but was spoken by 10 talkers, or ExpHi, in *List 2*.

were subdivided so that the familiarization was split into two experimental blocks. The term *block* refers to a familiarization phase followed by a testing phase. Four clusters were presented per block. The division of the experiment into two blocks was intended to limit the chance that children's attention spans were overly taxed.

Experimental Frequency and English Frequency were associated with each other such that, for a given block, all familiarization words in one condition of English Frequency were also in one condition of Experimental Frequency. With respect to the entire experiment, the two factors were combined so that there was a fully crossed 2 × 2 design, or four total conditions, illustrated in Table 3.5.

Table 3.5: The four within-subjects conditions in Experiment 1. The HiEng and LoEng conditions of the English Frequency factor were combined with the ExpHi and ExpLo conditions of the Experimental Frequency factor to create a 2 × 2 design.

	English Frequency	
Experimental Frequency	**HiEng**-*ExpHi*	**LoEng**-*ExpHi*
	HiEng-*ExpLo*	**LoEng**-*ExpLo*

Neighborhood density Several experiments have shown that phonotactic probability is highly correlated with neighborhood density, but that both have dissociable effects in various psycholinguistic tasks (Vitevitch & Luce, 1998; Bailey & Hahn, 2001). Because the latter was of interest here, it was important to establish that neighborhood density did not vary across the word sets or with respect to the English Frequency factor.

Neighborhoods for the word sets were determined using the one-phoneme edit distance metric[9] (Luce, 1986) in the Washington University Neighborhood Database (http://128.252.27.56/Neighborhood/Home.asp). None of the familiarization or test words had lexical neighbors, so neighborhood density was effectively the same for all words, in all word sets, and across both English Frequency conditions.

[9] An *edit distance* is the substitution, addition, or subtraction of one phoneme in a word that results in a new word. For example, the substitution of a /p/ for the /k/ in *cat* results in the new word *pat*; the addition of /s/ at the beginning of the word results in *scat*; the subtraction of the /k/ results in *at*. An edit distance of one refers to the fact that only one change is being made.

However, another experiment by Hollich et al. (2002) showed that 17-month-olds learned words faster in an artificially created dense neighborhood compared to a sparse one. A number of fairly similar words were used in the present experiment, but it is unlikely that an artificial neighborhood of the type seen in Hollich et al.'s work would have been created here. First, this is because the word presentation was divided into two blocks, and only four words appeared in each block. The words that shared a word frame, such as /mæfpəm/ and /mæstəm/, appeared in different blocks, so they were unlikely to have colluded to create some kind of neighborhood effect. Second, the differences in clusters, as well as the differences in word-initial CV sequences, prevented any of the words from having an edit distance of one or even two. Therefore, online neighborhood effects were not expected to occur in this experiment.

Familiarization Word Lists

The two familiarization lists were further balanced by reversing the blocks to create two additional lists. Familiarization words presented in the first block in one list were also presented in the second block in the other list. For example, in List 1A the HiEng familiarization word /mæstəm/ was associated with 10 talkers and appeared in the first block, whereas /fospəm/ was associated with just 1 talker and appeared in the second block. This list was reversed to create List 1B, in which /fospəm/ (still spoken by 1 talker) appeared in the first block and /mæstəm/ (still spoken by 10 talkers) appeared in the second block. The full set of familiarization lists (1A, 1B, 2A, and 2B) is given in Appendix A.

Test Words

The test word set (cf. Table 3.3) consisted of eight new words whose medial consonant clusters were the same as the medial clusters in the familiarization word set. These words were recorded in the same manner described above, but a novel talker (female, native speaker of American English with a Western US accent) was used. The test words were also recorded so that important phonetic cues to each consonant in the cluster (release of the first stop in the cluster, aspiration on the second stop) were clearly audible in the

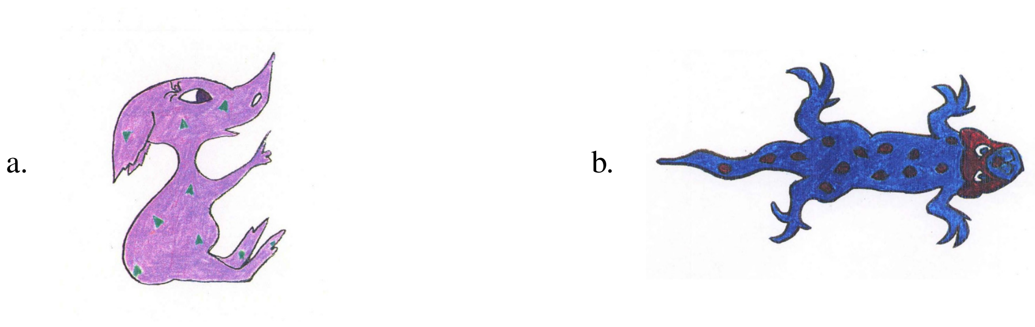


Figure 3.1: Two pictures that were used to create an experimental item. All familiarization and test words were assigned to unique animals. Animals were created by D. Ohala, S. Bourgeois, and A. Fountain.

test word token. The root-mean-square amplitudes of these tokens were equated with the familiarization words.

3.1.3 Procedure

Participants were brought in for a single experimental session. The experiment took place in a quiet, 9.5’×9.5’ room. Each experimental item consisted of a nonsense word accompanying by a hand-drawn picture of a make-believe animal (D. K. Ohala, 1999, cf. Figure 3.1). Children sat at a child-sized table with the computer screen and speakers on top. Children sat approximately 2’ away from the screen, and the speakers were placed directly behind the screen. The volume was set to a comfortable level. The presentation was controlled by Superlab 4.0 (www.superlab.com) software operating from an Apple Macintosh G4. The experimenter was seated next to the child and controlled the pace of the experiment from the laptop. The child’s parent(s) were in the experiment room or an adjacent room. They did not help the child with the experiment, but occasionally interceded to encourage the child if and when he or she became unhappy or distracted. If at any time the child became overly uncomfortable or indicated an unwillingness to continue, the experimenter stopped the experiment.

Before the first experimental block started, the experimenter explained to each child that he or she would play a game involving a series of “funny” or “make-believe” animals. For the familiarization phase, children were asked to “watch and listen” to the familiarization items. During the test phase, children heard the test words, which were associated with new make-believe animals. They were asked to “say these animals’ names back.” No emphasis was placed on speed, but results from the 10 Talkers studies (Gerken et al., 2006), which used a similar procedure, suggest that children generally repeated the words immediately. Before each phase of the experiment, the experimenter explained to the child what would happen next, and proceeded only when the child indicated his/her readiness. When the experiment ended, the child received a small gift.

Each familiarization phase contained four of the eight familiarization words. Two were presented 10 times (ExpHi) and two once (ExpLo), all in a random order, for a total of 22 familiarization tokens per block. During the test phase, each of the four test words was presented randomly in a series, for four total series.

3.1.4 Analysis

In the test phase of the experiments, each child was asked to produce four tokens of each word, for a total of four possible data points per item. Two dependent measures were collected for each production: accuracy on the two medial consonants and the time from the end of the target word to the beginning of the child’s utterance, referred to hereafter as the production latency.

Accuracy For the accuracy measure, the children’s productions were transcribed and a score was tallied based on whether their production of the target cluster was entirely accurate (a score of ‘2’ for each consonant), whether one or more consonants was produced in error (a score of ‘1’ for each consonant in error), or whether a consonant was omitted (a score of ‘0’ for each omitted consonant). The maximum score for any given consonant cluster was ‘4’. A word produced without a medial cluster received a score of ‘0’.

A second transcriber independently transcribed all of the data, and discrepancies between the two sets of transcriptions were eliminated in a second pass, during which both transcribers listened to items for which their transcriptions diverged. If they then agreed on a transcription, this transcription was entered into the accuracy analysis so that 100% of the analyzed data was transcribed identically by two transcribers. If the transcribers were unable to agree on a transcription, however, the item was removed from the analysis. Less than 2% of the data was removed in this way with no more than three items removed per participant.

Production Latencies Production latencies were measured from the offset of the target word to the onset of the child's production. This measurement was sensitive to the Experimental Frequency manipulation in the 10 Talkers experiments, so it was also used here. There is some reason to believe that it is not a particularly sensitive measure, however. Munson, Swenson, and Manthei (2005) measured a number of phonetic qualities of children's productions, including production latencies (or offset-to-onset times), and onset-to-onset times. The latter measurement was found to be more sensitive to the phonotactic characteristics of words. I follow Gerken et al. (2006) in using production latencies, but non-significant effects may reflect a lack of sensitivity inherent to this measurement.

All of the test words ended in nasals, consonants which often have low amplitudes and are difficult to see on a spectrogram. For this reason, production latencies were not determined by eyeballing the time from the offset of the target to the onset of the child's production. Instead, because the durations of the test words are known, a precise method can be used for finding the latencies using the target word onsets. Onsets are relatively easy to find on a spectrogram, so from the onset of the target, the word length was added to find a point corresponding to the target offset. From there, the distance to the onset of the child's production was determined. This procedure was implemented using Praat software (www.praat.org).

Production latencies were also analyzed independently by a second data coder. The similarity between the two sets of measurements was quite high. The two coders were within 50 ms of one another for 77.5% of their measurements and within 100 ms for 93.2%. If productions were inaudible or if children stopped and attempted their productions a second time, these productions were removed from the analysis. Approximately 5% of the data was removed for this reason.

The average production latency and accuracy scores across a given repetition are given in Figure 3.2. An average score was obtained for the eight tokens of each repetition for each subject (eight 1^{st} repetitions, four from each experimental block; eight 2^{nd} repetitions, etc.), then those averages were collapsed to create the figure. The most precipitous change from one set of repetitions to the next is from the first to the second repetition in the production latency data. Participants often took several seconds to respond to the first few items. Due to a high degree of variation in the production latencies for the first repetitions, they were removed from the analysis. Although there were also differences between the second, third, and fourth repetitions, these differences were much smaller. With respect to the accuracy averages, there was not a precipitous change in accuracy scores from one repetition to the next, so all four repetitions were included in the accuracy analysis.

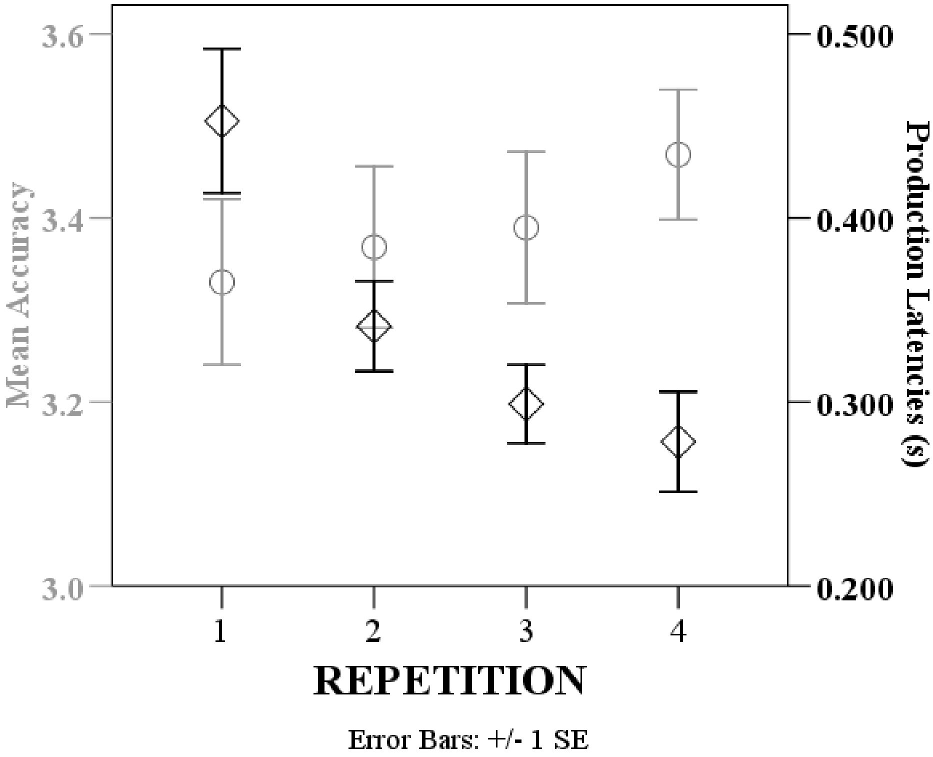


Figure 3.2: Average accuracy and production latency scores for each of the four repetitions collected in Experiment 1. The accuracy averages are represented by open circles and are shaded gray. The black diamonds are the averages for the production latencies. The scale for the accuracy averages is on the left, the scale for the production latency averages is on the right.

3.1.5 Results

Table 3.6 gives mean accuracy and production latency scores for the remaining data for each of the four conditions in the experiment (HiEng+ExpHi, HiEng+ExpLo, LoEng+ExpHi, LoEng+ExpLo, cf. Table 3.5), as well standard deviations. Accuracy results were collapsed across all repetitions, items, and clusters to give a unique data point for each of the four conditions for each participant. Because the first repetitions were removed from the production latency analysis, the production latency data were collapsed across the second, third, and fourth repetitions only. Comparisons of the four conditions for both dependent measures were made using a 2 × 2 within-subjects ANOVA.

Table 3.6: Mean accuracy (out of 4) and production latency (in seconds) scores for Experiment 1. Standard deviations are given in parentheses.

Accuracy				
	HiEng *ExpHi*	*HiEng* *ExpLo*	*LoEng* *ExpHi*	*LoEng* *ExpLo*
Mean	3.533	3.519	3.268	3.182
(Std Dev)	(.452)	(.484)	(.564)	(.502)

Production Latency				
	HiEng *ExpHi*	*HiEng* *ExpLo*	*LoEng* *ExpHi*	*LoEng* *ExpLo*
Mean	.318	.296	.298	.315
(Std Dev)	(.112)	(.115)	(.079)	(.117)

For the analysis of accuracy, there was a significant effect of English Frequency, HiEng words were produced more accurately than LoEng words (F (1,15) = 6.380, $p < .05$, $\eta_p^2 = .298$), but not of Experimental Frequency (F (1,15) = 0.103, $p = .753$, $\eta_p^2 = .007$), and no English Frequency × Experimental Frequency interaction (F (1,15) = .178, $p = .679$, $\eta_p^2 = .012$). Figure 3.3 provides graphs of the results for the accuracy analysis.

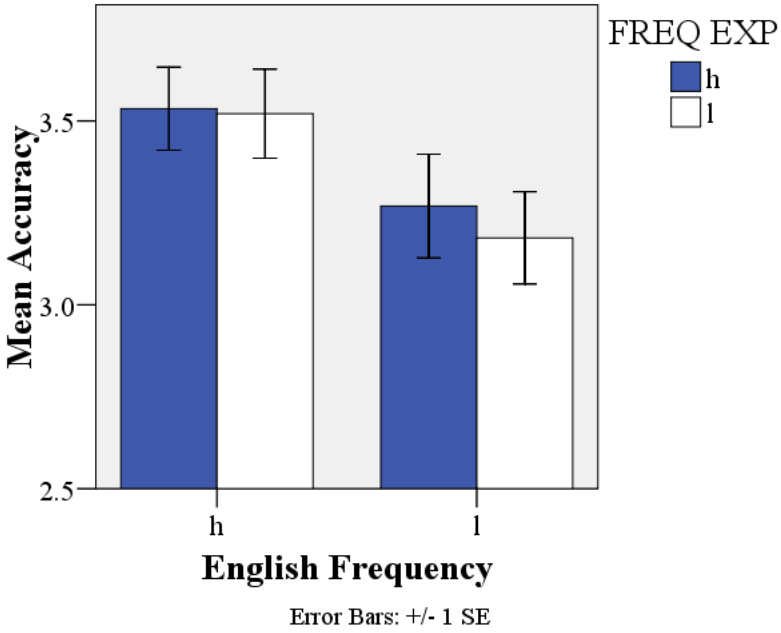


Figure 3.3: Bar graphs of the accuracy results from Experiment 1. On the ordinate is the mean accuracy for the medial consonant clusters, with a score of '4' being accurate production of both consonants. The bars are shaded according to Experimental Frequency. Blue bars represent the average accuracy for the ExpHi clusters, white bars represent the ExpLo clusters. Bars are grouped by English Frequency. Results for the HiEng clusters are on the left, results for the LoEng clusters are on the right.

In the analysis of production latencies, there was no significant effect of English Frequency (F (1,15) = .005, p = .945, η_p^2 = .000) or Experimental Frequency (F (1,15) = .010, p = .922, η_p^2 = .001), and no English Frequency × Experimental Frequency interaction (F (1,15) = 2.612, p = .127, η_p^2 = .148). These results suggest that perceptual familiarization with multiple tokens of a phonotactic sequence did not lead to generalization in the production of a new word, even when token variability is present. Figure 3.4 provides graphs of the results for the production latency analysis.

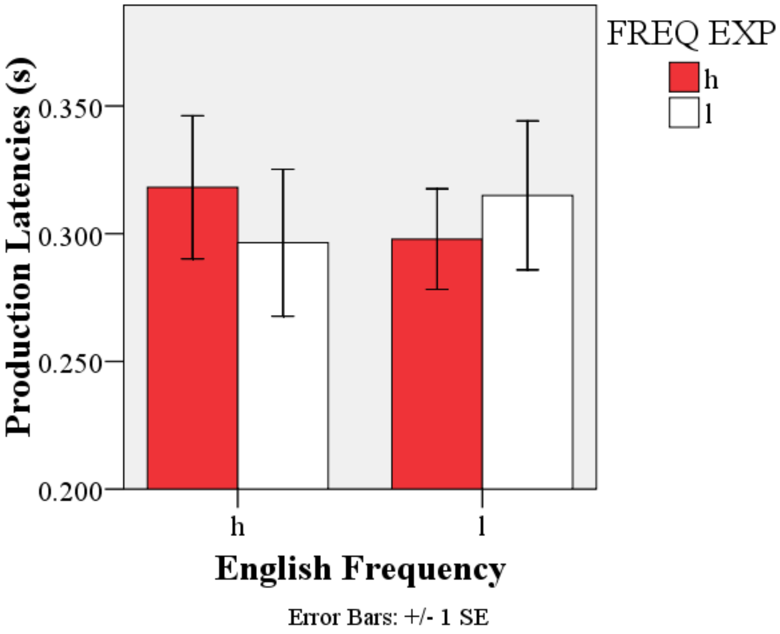


Figure 3.4: Bar graphs of the production latency results from Experiment 1. On the ordinate is the time from the end of the target word to the onset of the child's production in seconds. The bars are shaded according to Experimental Frequency. Red bars represent the average production latencies for the ExpHi clusters, white bars represent the ExpLo clusters. Bars are clustered by English Frequency. Results for the HiEng clusters are on the left, results for the LoEng clusters are on the right.

Distribution of the Data

In an effort to better understand what factors might have contributed to the null effect, the distribution of the data was analyzed. Histograms of the accuracy scores for the HiEng and LoEng clusters are given in Figure 3.5. Some effects of the Experimental Frequency manipulation may have been hidden by a ceiling effect, particularly for accuracy measurement of the HiEng clusters. This concern will become more relevant for subsequent experiments, but an analysis of normality is presented here, as well. To determine whether there was a ceiling effect present in the accuracy measurement, average accuracy scores for each subject for the HiEng and LoEng clusters were analyzed using a Kolmogorov-Smirnov test for normality. It should be noted that a Kolmogorov-Smirnov test for normality does not test how skewed a data set is or show which direction the data is skewed. For this reason, I interpret the tests, in part, based on visual inspection of the histograms. There was a non-significant trend for the HiEng clusters (*D* (*df* = 16) = .211, *p* = .054), which, based on the histogram in Figure 3.5b, appears to have resulted from a reduced spread above the mean, but no significant difference from a normal distribution for the LoEng clusters (*D* (*df* = 16) = .076, $p \geq .200$). Thus, it is possible that effects of the Experimental Frequency manipulation were hidden for the HiEng clusters[10].

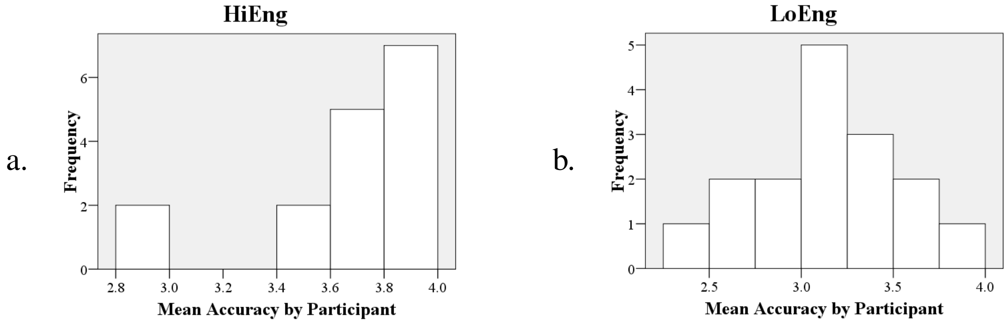


Figure 3.5: Histograms of the accuracy measurements for the HiEng and LoEng conditions in Experiment 1. Accuracy scores for the HiEng clusters are given in Panel a and scores for the LoEng clusters are given in Panel b.

[10] Although this procedure is not ideal, I know of no statistical means for testing skew.

3.1.6 Discussion

Corroborating numerous past studies (Beckman & Edwards, 1999; Edwards et al., 2004; Munson, 2001; Zamuner et al., 2004), the children in Experiment 1 produced consonant clusters with high phonotactic probabilities more accurately than clusters with low phonotactic probabilities. However, children's prodctions were not influenced by a manipulation of Experimental Frequency. That is, children were not faster or more accurate when producing either HiEng or LoEng clusters after having heard those clusters in a related word spoken by 10 different talkers. The results suggest that children were unable to apply knowledge of a phonotactic sequence present in a familiarization word to their productions of the same sequence in a novel test word, although there remains the possibility that some effects of Experimental Frequency were hidden, at least for the HiEng clusters, by a ceiling effect. The ceiling effect will be covered in greater detail in subsequent chapters.

This possibility aside, the results suggest that token variation did not generalize. This finding is in contradiction to the predictions made by the exemplar-based model of Bailey and Hahn (2001) and Goldinger (1998). In Bailey and Hahn's model, token frequency is counted by the model and influences the the model's ability to generalize about phonotactic sequences. That fact, combined with Goldinger's proposal that words be linked by common phonetic properties, suggests that token-based generalization should be possible. When children heard a cluster produced by 10 talkers in the ExpHi condition, the expectation was that they would be more accurate when producing that cluster in a new word compared to when they had heard just a single token of that cluster in the ExpLo condition. Children were neither faster nor more accurate when producing the ExpHi words, however, providing no evidence for the predicted generalization in either measure. The results pose a problem for exemplar models, as they suggest that token frequency of a sublexical pattern does not influence how sensitive children are to that pattern.

The problem for exemplar models becomes more clear when we compare the present results to the results from the 10 Talkers experiment (Gerken et al., 2006). In that study, token variability in the familiarization phase did lead to increased production speed and accuracy for productions of the same word. However, the exemplar architecture that seemed apt for explaining the 10 Talkers results is also what motivated the prediction that familiarization with multiple talkers should generalize: talker-tokens should be linked by the phonetic properties that each token has in common, and those same links should be activated by a related word. The null result suggests that this view of the representation of talker variability is flawed. Rather, the results of the 10 Talkers experiment compared to the present results suggest that talker variability allows listeners to develop clear representations of a familiarization word's shape, or its broader properties, but does not provide them with a representation of the word's sublexical content that is relevant beyond the scope of the word.

With respect to perceptual learning, the results also conflict with the conclusions drawn in Gerken et al. (2006) about learning phonotactics. Perceptual learning from multiple talkers had a strong influence on productions of the same word, but it does not appear to allow for generalizations about the word's sublexical content, or phonotactic probabilities, at least based on token variability alone. It therefore remains unclear whether the main effect of English Frequency found here, or similar effects found elsewhere (Beckman & Edwards, 1999; Munson, 2001; Edwards et al., 2004) should be interpreted as reflecting perceptual learning, or articulatory practice, or both.

It should be noted that the results are not entirely inconsistent with the conclusions reached by Bailey and Hahn (2001), who show that an exemplar-based model of neighborhood density, the Generalized Neighborhood Model, provides an incomplete account of adult wordlikeness judgments, and that phonotactic probabilities account for a significant portion of people's judgments, as well. Bailey and Hahn argue that more is to be gained from increasing the complexity of neighborhood models, and neighorhood activation may eventually subsume phonotactic frequency effects in subsequent models. One possible form of increased complexity, increased phonetic detail, is precisely what is present in Goldinger's (1998) and Johnson's (1997) models, but is also what appears to have failed here. The phonetic details necessary to connect word tokens from multiple talkers should have allowed children to learn the internal structure of the familiarization words, leading them to generalize the target phonotactic sequences to their productions of the test words. This did not happen, however, indicating that an increase in the complexity of neighborhood models may not be as advantageous as Bailey and Hahn suggest. In fact, it seems reasonable to attribute some of the success of the the Generalized Neighborhood Model to the fact that it was largely insensitive to phonetic-level exemplar dynamics.

Also worth noting is the fact that the criticism being raised against the exemplar model is based on a null result that was predicted to be significant. This is a somewhat limited form of argumentation when criticizing a theory or model. However, the possibility that additional experiments may reveal generalization effects does pose a serious problem for exemplar models. If children are able to generalize in subsequent experiments, the exemplar model—which predicts significant effects in all experiments—is at loss to explain why generalization was found in some experiments but not in others. This possibility, combined with the fact that an experiment could not be designed in which the exemplar model should predict a *null* result, provides a much stronger platform from which exemplar models can be criticized.

Although the results raise questions about the ability of exemplar models to capture phonotactic generalization effects, the null result in Experiment 1 is completely in line with predictions made by Pierrehumbert's hybrid model and Albright's Minimal Generalization Learner. These models claim that type frequency, rather than token variability, is the critical ingredient for learning phonotactics. To test whether type frequency would allow children to generalize a phonotactic sequence, two additional experiments were conducted in which the number of familiarization words containing a target cluster varied. This type frequency manipulation was combined with token variability: each familiarization word was spoken by four different talkers. The results of the first word-types + word-tokens experiment, Experiment 2A, are reported next.

CHAPTER 4

EXPERIMENT 2A: PHONOTACTIC GENERALIZATION FROM MULTIPLE WORD-TYPES AND MULTIPLE TALKER-TOKENS

As discussed in Chapter 2, past research on perceptual learning has shown that children are faster and more accurate at producing a word if they first hear that word spoken by 10 Talkers (Gerken et al., 2006). Experiment 1 asked whether this “10 Talker effect” would generalize to productions of the same phonotactic sequence in a new word. Children participating in that experiment were familiarized with 10 tokens of a CVCCVC nonsense word, and were then asked at test time to produce a new word containing the same medial cluster. There were no significant effects of familiarization in that experiment, however. One possible reason for the null effect is that type frequency, rather than token frequency, is essential for a phonotactic pattern to be learned (Albright, 2007; Bybee, 1995; Pierrehumbert, 2003b).

In Experiment 2A, discussed here, four-year-old children were tested on their ability to generalize a phonotactic sequence based on token variability and type frequency. As in Experiment 1, the experimentally defined sequences were eight word-medial consonant clusters in CVCCVC nonsense words. Four of the clusters had high phonotactic probabilities (HiEng) and four had low or zero probabilities (LoEng). During a perceptual familiarization phase, children heard these clusters in either a single nonsense word (ExpLo) or three nonsense words (ExpHi). Children also heard four tokens of each nonsense word, with each token produced by four different talkers. At test time, children were asked to produce the same medial clusters in new words. The results were analyzed for accuracy and speed (production latency).

Recall that, in Chapter 2, the predictions of three models were contrasted based on how type frequency and token variability might play out in phonotactic learning. A summary of these predictions is repeated in Table 4.1 below. Albright’s (2007) Minimal Generalization Learner and Pierrehumbert’s (2003b) hybrid model both predict that type frequency is critical for generalization, so they both predict significant results of Experimental Frequency in Experiment 2A. However, exemplar models also allow for generalization based on type frequency to the extent that it is consistent with token frequency. That is precisely the case here, Therefore, all three models predict improved productions for the ExpHi clusters versus the ExpLo clusters, although no model's predictions are specific to either speed, or accuracy, or both.

Table 4.1: Predictions made by the three models—Bailey and Hahn's exemplar model, Pierrehumbert's hybrid model, and Albright's MGL—for the word-types + word-tokens experiment. *Gen* is given in cells where a model (row) predicts a significant effect of the Experimental Frequency manipulation for a given experiment (column). *No Gen* refers to a predicted null effect. A prediction that is unique to a particular model is boxed and in bold.

	EXP 1	EXP 2A	EXP 2B	EXP 3
	Tokens	Types + Tokens	Types + Tokens	Types
EXEMPLAR MODEL Bailey and Hahn (2001)	**Gen**	Gen	Gen	Gen
HYBRID MODEL Pierrehumbert (2003b)	No Gen	Gen	Gen	**No Gen**
MINIMAL GENERALIZATION LEARNER Albright (2007)	No Gen	Gen	Gen	Gen

4.1 Method

4.1.1 Participants

Twenty-two children were recruited for this study. All were between the ages of 4;0 and 4;3 years of age (mean = 4;1.24). As in Experiment 1, they were recruited using a database of local birth announcements and adoptions in the Tucson metropolitan area. All of the children were native English speakers, as reported by their parent/caregiver in a short questionnaire, and had minimal exposure to another language (never more than 10 hours, average was 3.5 hours, mode was 0 hours, six children had some exposure to Spanish, one child to American Sign Language, and one to Chinese). Nineteen children were reported to have no personal or family history of any of the following: early intervention services, ear infections in the month prior to their participation, congenital hearing loss, congenital language delay, or speech or language therapy. Two children with speech delays and one child with a family history of speech disorders were removed from the analysis. Two more children were removed due to an equipment malfunction, and one child did not complete the study. Results from the remaining 16 children, 8 girls and 8 boys, are discussed below.

4.1.2 Materials

As in Experiment 1, children were familiarized with one set of words and then asked to repeat words from another set. The familiarization words (ExpHi and ExpLo word sets)[1] are given in Table 4.2. Of interest were the medial clusters and whether perceptual familiarization with words containing these clusters would influence productions of the clusters in novel test words. Some of the clusters were appeared in multiple familiarization words and others in a single word. This the purpose of this manipulation was to determine whether the number of words that instantiated a cluster, or the cluster's type frequency, influenced generalization.

Assuming that children could generalize phonotactic sequences common across multiple words, it was important to create words that supported the desired generalizations while avoiding unwanted generalizations. To this end, the words were assembled so that their medial consonant clusters were the only phonotactic sequences that were consistently present in the familiarization words. Specifically, all of the words that children heard at a given time had to be sufficiently different so that the only obvious generaliza-

[1] Once again, *set* refers to where the words appeared in the experiment (cf. Section 3.1.2 in Chapter 2). In Experiment 1, there was a perceptual familiarization word set and a test word set. In Experiment 2A, there was an ExpHi set, with three words per clusters, an ExpLo set, with one word per cluster, and a test word set, also with one words per cluster. The term *list* refers to the assignment of clusters to the Experimental Frequency conditions ExpHi and ExpLo in different lists. In Experiment 2A, words for the cluster /sp/ in the *ExpHi set* are /kɛspəs/, /tuspən/, and dɪspək/. As shown in Table 4.2, these words appear in *List 2*. In contrast, the *ExpLo set* contains only one /sp/ word, /fospəm/, which appears in *List 1*. The full set of lists is given in Appendix A.

Table 4.2: Familiarization words used in Experiment 2A. The words are arranged according to their distribution to lists and experimental blocks. For example, the HiEng clusters /mp/ and /kt/ and the LoEng clusters /mk/ and /pk/ are ExpHi in List 1 and so are present in three experimental words each. The HiEng ExpHi clusters appear in Block 1, while the LoEng ExpHi clusters appear in Block 2. In List 2 these clusteres are ExpLo and are present in just a single word, whereas the clusters /sp/, /st/, /ʃp/, and /fp/ are all ExpHi. Words are written out in IPA.

	Eng Freq	**List 1**		**List 2**		**Exp Freq**
Block One	**High English**	**mp** dompət sæmpəs gumpət	**kt** lɛktəf maʊktəs saktəf	**sp** kɛspəs tuspən dɪspək	**st** sʌstəp lostən gɪstək	*ExpHi*
	Low English	**ʃp** foʃpəm	**fp** mæfpəm	**mk** fæmkɪm	**pk** bopkəm	*ExpLo*
Block Two	**High English**	**sp** fospəm	**st** mæstəm	**mp** fæmpɪm	**kt** boktəm	*ExpLo*
	Low English	**mk** domkət sæmkəs gumkət	**pk** lɛpkəf maʊpkəs sapkəf	**ʃp** kɛʃpəs tuʃpən dɪʃpək	**fp** sʌfpət lofpən gɪfpək	*ExpHi*

tion across words were the medial clusters common to the ExpHi words[2]. The experimental procedure was divided into two blocks of familiarization followed by testing, with four of the eight clusters being presented in a block, so it was important that no unwanted generalizations existed among any of the words used for a given block.

As can be seen in Table 4.2, no word-initial consonant appeared more than twice in a given block, nor did the stressed vowel of the first syllable. Furthermore, the ExpLo words started exclusively with labial consonants, whereas only two ExpHi words started with a labial consonant, /maʊktəs/ and /maʊpkəs/. Although this design limited the number of possible generalizations present in the familiarization words, it had the unintended consequence of creating a place of articulation confound. Because the ExpLo

[2] The second vowel in the word could be added to this generalization, although it is consistent for all familiarization and test words, making it redundant and uninformative when interpreting a generalization effect.

words *always* started with a labial consonant, it is possible that an effect of Experimental Frequency could be attributed to the place of articulation of the initial consonant, rather than the type frequency of the medial cluster. This is an unlikely scenario, but I will return to this confound in the discussion. Note that this confound was removed in Experiment 2B (discussed in Chapter 5).

The word sets were kept small so that the familiarization phase was relatively short and a reasonable number of repetitions could be collected for each test word. This design, as in the previous experiment, does not contain enough words for a proper by-items analysis, but such an analysis is arguably inappropriate for this experiment (cf. Section 3.1.2 of Chapter 3) because the same words are used in both conditions of the Experimental Frequency factor, and the word frame was used for both a HiEng and a LoEng cluster, so that the English Frequency conditions also contained highly similar words. Thus, the words in this experiment should not be considered as a randomly distributed variable, obviating any results obtained from a by-items analysis. See section 3.1.2 for further discussion of this issue.

English Frequency

As in Experiment 1, the phonotactic sequences common to both the familiarization and the test words (Munson, 2001) were the High English Frequency (HiEng) clusters /kt/, /mp/, /sp/, and /st/, and the Low English Frequency (LoEng) clusters /pk/, /mk/, /ʃp/, and /fp/. Vitevitch's Phonotactic Probability Calculator was again used to determine the phone and biphone sums, or the sums of all component phones or biphones present in a word, respectively, for each group of words. Sums for the three sets of words—the set with three words per cluster (ExpHi), the set with one word per cluster (ExpLo), and the test word set—are given in Table 4.3.

Also as in Experiment 1, it was important to establish that the word sets were balanced. Using the phone and biphone sums as dependent variables, an ANOVA with English Frequency and Set as factors was conducted (Set refers to whether the words were in the ExpHi, ExpLo, or test group). For the analysis of the phone sums, there was a significant effect of English Frequency (F (1,18) = 12.574, $p < .01$) but not of Set (F (2,18) = 1.888,

Table 4.3: Sums of phone and biphone probabilities for the familiarization and test words used in Experiment 2A.

		Phone Sums			**Biphone Sums**		
English Freq	*Cluster*	*ExpHi*	*ExpLo*	*Test*	*ExpHi*	*ExpLo*	*Test*
High	**kt**	0.3605	0.4200	0.3645	0.0259	0.0319	0.0198
	mp	0.2918	0.3847	0.3633	0.0195	0.0342	0.0268
	sp	0.3280	0.3430	0.3843	0.0325	0.0317	0.0361
	st	0.4219	0.3861	0.4127	0.0577	0.0505	0.0470
Low	**pk**	0.2970	0.3565	0.3010	0.0171	0.0238	0.0110
	mk	0.2978	0.3907	0.3694	0.0095	0.0230	0.0156
	Sp	0.2569	0.2719	0.3132	0.0229	0.0214	0.0201
	fp	0.3096	0.2738	0.3292	0.0272	0.0232	0.0185

p = .180), and no interaction (F (2,18) = 0.022, p = .978). For the analysis of the biphone sums, there was also a significant effect of English Frequency (F (1,18) = 14.112, $p < .01$) but not of Set (F (2,18) = 0.666, p = .526), and no interaction (F (2,18) = 0.020, p = .980). Because the words were balanced such that the same word frame was used for a HiEng and a LoEng cluster, the significant differences in English Frequency can be attributed to the clusters[3]. The nonsignificant effects of Set and the lack of interactions suggest that the different word sets were balanced. In other words, the word frames used for the ExpHi words did not differ from the word frames used for the ExpLo or test words.

Syllable structure As discussed in Section 3.1.2 of the previous chapter, there are reasons to think that the English Frequency factor reflects differences in syllable structure as well as phonotactic probabilities. This is because the HiEng clusters /sp/ and /st/ are allowable as onsets in English and all four HiEng clusters are allowable as codas, but the four LoEng clusters are virtually unattested as either onsets or codas in English. Section 3.1.2 also includes several arguments for interpreting the English Frequency factor

[3] The individual biphone scores adjacent to the medial cluster were different for the HiEng and LoEng conditions. This is to be expected, given that different phonemes were used in the HiEng clusters compared to the LoEng clusters. Although this difference may have had some bearing on the results for the English Frequency factor, they should not be interpreted with respect to the Experimental Frequency factor, in which items appeared in both the ExpHi and ExpLo conditions.

as a measure phonotactic probability, including the fact that similar studies have used regression analyses and found significant contributions of the phonotactic probabilities of consonant clusters to children's production accuracy, regardless of syllable structure.

Recall, too, that the primary variable of interest is not English Frequency, or the intrinsic frequency of a cluster, but the Experimental Frequency, or the number of times that children hear a cluster during the familiarization phase of the experiment. The Experimental Frequency variable is what informs us about how children learn phonotactics and about which qualities of the input lead to phonotactic generalization. Therefore, the English Frequency factor and its interpretation is only of modest interest. In Experiment 2A, I will again refer to the English Frequency factor as reflecting phonotactic probabilities, consistent with previous studies, but I will also return to the issue of syllabification in the discussion.

Experimental Frequency

As before, Experimental Frequency was a manipulation of the number of words associated with a given consonant cluster. However, in Experiment 2A Experimental Frequency was a manipulation of type frequency: clusters in the Experimental Frequency high (ExpHi) condition were heard in three different words, clusters in the Experimental Frequency low (ExpLo) condition were heard in a single word. In order to have all eight clusters appear in both the ExpHi and ExpLo conditions, two word lists were created, and a cluster appeared as ExpHi in one list and ExpLo in the other. The division of words according to the ExpHi and ExpLo conditions and the two lists is given in Table 4.2. For example the cluster /sp/ was ExpLo in Block 2 of List 1, and appeared in the word /fospəm/. In Block 1 of List 2, /sp/ was ExpHi and was associated with three words: /kɛspəs/, /tuspən/, and /dɪspək/.

Each word was also presented with a baseline level of talker variability. Four tokens of each familiarization word were used, with a different talker producing each token. Four women, all native American English speakers of a Western US dialect, were recorded producing each word. A single token from each talker, for each familiarization word, was chosen to be played during the familiarization phase of the experiment. The same recording setup described for Experiment 1 was used, and care was again taken to ensure that each token had clear phonetic cues for the consonant cluster, and that stress was on the first syllable. Tokens were extracted with 100 ms of silence on either side, and the root-mean-square amplitude of all tokens were equated.

A given block of familiarization included 8 words (3 words each for the ExpHi clusters and 1 word each for the ExpLo clusters), and each word was produced by 4 talkers, so children heard the ExpHi clusters 12 times each and the ExpLo clusters 4 times each. This deviates from the 10 tokens versus 1 token setup of Experiment 1; however, the change was necessitated by the addition of multiple words for each cluster. Given that type frequency and not token variability was the critical variable, the change to 12 tokens versus 4 tokens was inevitable. Although this design results in a greater degree of variability—whether in terms of talkers, or words, or both—compared to Experiment 1, the ratio of variability between ExpHi and ExpLo conditions was actually greater in the previous experiment (10:1 versus 12:4). The variability ratio should therefore have favored generalization in Experiment 1, so it is unlikely that any generalization effects found here are due to the greater amount of variability. Furthermore, an advantage of this design is that, given that the ExpLo words were heard by multiple talkers, this condition is quite similar to the ExpHi condition of Experiment 1. Therefore, the design allowed for a comparison of type frequency combined with token variability compared to token variability alone, rather than just comparing one word to three. This design therefore provides a strong test of type frequency, as children's performance in the ExpHi condition must be above and beyond any benefits that are due to token variability.

Neighborhood density Due to the high correlation between phonotactic probability and neighborhood density (Vitevitch & Luce, 1998; Bailey & Hahn, 2001), it was important to establish that the neighborhood density of the familiarization and test words did not differ in terms of neighborhood density. Neighborhoods for the word sets were determined by the one-phoneme edit distance metric (Luce, 1986) using the Washington University Neighborhood Database (http://128.252.27.56/Neighborhood/Home.asp). None of the familiarization or test words had lexical neighbors, so neighborhood density was effectively the same for all familiarization and test words.

With respect to the possibility of neighborhoods being created during the course of the experiment, or online (Hollich et al., 2002), the matched word pairs (for example, /mæfpəm/ and /mæstəm/) appeared in different blocks, and the differences in the word-initial consonants prevented any of the words from having an edit distance of one or even two, so it is unlikely that online neighborhoods could have been created (Hollich et al., 2002).

Familiarization Word Lists

The two word lists in Table 4.2 were rearranged in two new lists so that the clusters kt, mp, ʃp, and fp, which were heard in Block 1 in one list, were heard in Block 2 for another list. This created a total of four lists, each of which was heard by 4 of the 16 total participants. All four lists (1A, 1B, 2A, and 2B) are presented in Appendix A. Two girls and two boys were assigned to each list so that gender was balanced across lists.

Test Words

The test words are identical to those used in Experiment 1, and are repeated here in Table 4.4. The test set consisted of eight new words with the same medial consonant clusters that appear in the familiarization words.

The test words were recorded in the same manner described above for the familiarization words, but a different talker (female, native speaker of American English with a Western US accent) recorded them. As with the familiarization items, the talker was asked to produce the words so that important phonetic cues to each consonant in the cluster (release of the first stop in the cluster, aspiration on the second stop) were clearly audible. Representative tokens were chosen from the talker's recordings, and the root-mean-square amplitudes of these tokens were normalized.

The test words also differed with respect to their initial CVs compared to the words in the ExpHi and ExpLo word sets, which eliminated the possibility of unwanted generalizations present across familiarization and test words. The phone and biphone sums of the

Table 4.4: The test word set used in Experiment 2A. This set is identical to the set used in Experiment 1, listed in Table 3.2 These words are matched with the familiarization word sets for phone and biphone sums. Words are written out in IPA.

High English	**Low English**
sp /zaspən/	**ʃp** /zaʃpən/
st /neɪstən/	**fp** /neɪfpən/
mp /simpən/	**mk** /simkən/
kt /daktən/	**pk** /dapkən/

test words were matched with the sums of the familiarization word sets, as is evidence by the lack of significant differences found between the word sets in the analysis described in the description of the English Frequency factor.

4.1.3 Procedure

The procedure was the same as in Experiment 1, except that there were three items (nonsense word + make-believe animal pairings) for each ExpHi cluster and one item for each ExpLo cluster.

As in Experiment 1, each experimental session consisted of two blocks of familiarization followed by testing. With four clusters per block, children heard a total of 32 familiarization tokens (2 ExpHi clusters × 3 words × 4 talker-tokens = 24 tokens, 2 ExpLo clusters × 1 word × 4 talker-tokens = 8 tokens). During the test phase, four repetitions of each test word were collected across four randomized series.

4.1.4 Analysis

The same dependent measures used in Experiments 1 and 2A were collected here: accuracy on the two medial consonants and production latencies. The same system for computing accuracy scores and determining latencies was also used here. Accuracy scores were based on four possible points for a cluster, or the combination of two points per component consonant. A correct production of a component consonant was scored as '2', in incorrect production was scored as '1', and a missing consonant was scored as '0'. Two data coders independently transcribed the data and worked together to eliminate discrepancies. As in Experiment 1, both coders went through and transcribed items where they were not in agreement about the transcription of the medial cluster. If no agreement between the coders could be reached, the item was removed from the analysis. Less than 2% of the data was removed due to a lack of agreement, and never more than three items were removed per participant.

Production latencies were determined by first finding the length of the target word in the sound file, then finding the time from the end of the target to the beginning of the child's own production. Measurements were made using Praat software (www.praat.org) and entered into a spreadsheet for analysis. Reliability for the production latency analysis was quite high between the two independent coders. Approximately 89.0% of the measurements made by both coders were within 50 ms of each other, and 97.2% were within 100 ms. Approximately 4% of the data was removed because children attempted their productions two or more times.

As in Experiment 1, the first repetition of each word was removed from the production latency analysis due to the high level of variability in those productions. Figure 4.1 gives the accuracy and production latency scores averaged across the eight repetitions obtained from each subject. In this experiment, the production latencies for the first repetition were again much longer than for any of the other repetitions. For the accuracy analysis, results across the four repetitions were relatively consistent, so all of these data were included in the analysis.

Finally, one HiEng item, /zaspən/ was removed from the analysis because of an experimental setup error.

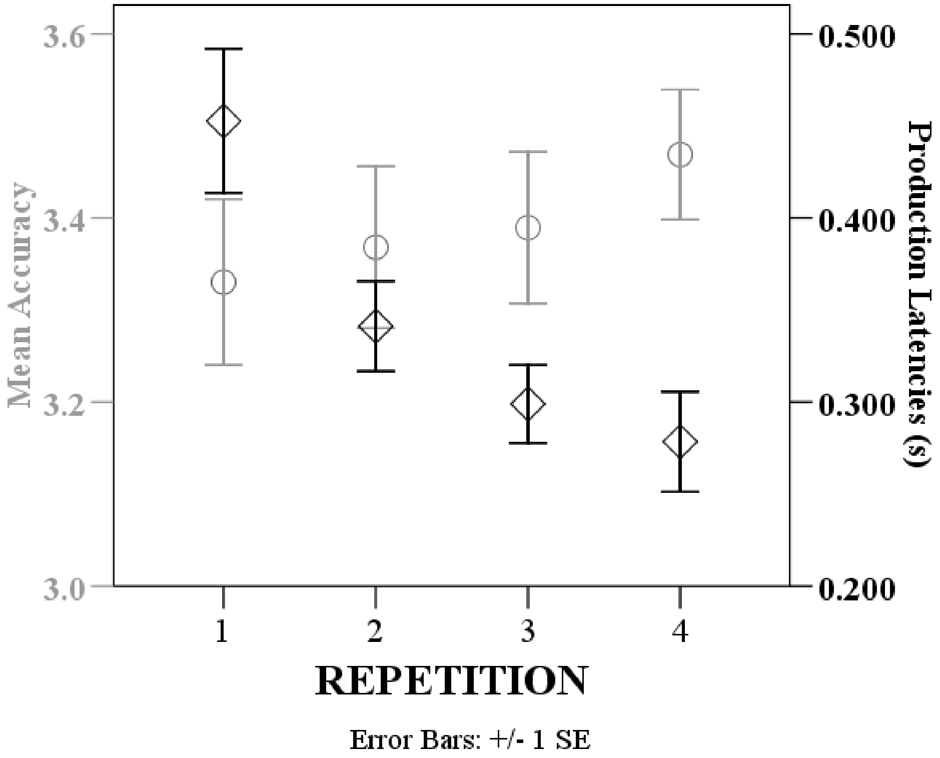


Figure 4.1: Average accuracy and production latency scores for repetitions 1-4 from Experiment 2A. The accuracy averages are represented by open circles and are shaded gray. The black diamonds are the averages for the production latencies. The scale for the accuracy averages is on the left, the scale for the production latency averages is on the right.

4.1.5 Results

Table 4.5 gives mean accuracy and production latency scores and their standard deviations for each of the four experimental conditions (HiEng+ExpHi, HiEng+ExpLo, LoEng+ExpHi, LoEng+ExpLo). Separate 2×2 within-subjects ANOVAs were then conducted for each dependent measure.

Table 4.5: Mean accuracy (out of 4) and production latency (in seconds) scores for Experiment 2A. Standard deviations are given in parentheses.

Accuracy				
	HiEng	*HiEng*	*LoEng*	*LoEng*
	ExpHi	*ExpLo*	*ExpHi*	*ExpLo*
Mean	3.282	3.307	3.136	2.745
(Std Dev)	(.493)	(.526)	(.503)	(.524)

Production Latencies				
	HiEng	*HiEng*	*LoEng*	*LoEng*
	ExpHi	*ExpLo*	*ExpHi*	*ExpLo*
Mean	.275	.275	.268	.275
(Std Dev)	(.103)	(.103)	(.097)	(.103)

For the accuracy analysis, there was a significant effect of English Frequency (F (1,15) = 9.824, $p < .01$, $\eta_p^2 = .396$) resulting from greater accuracy scores for HiEng words compared to LoEng words. There was not an effect of Experimental Frequency (F (1,15) = 2.094, $p = .168$, $\eta_p^2 = .122$), but there was a significant English Frequency $\times$ Experimental Frequency interaction (F (1,15) = 5.154, $p < .05$, $\eta_p^2 = .256$). Comparing the simple effects, a significant effect of Experimental Frequency for the LoEng clusters (F (1,15) = 9.787, $p < .01$, $\eta_p^2 = .395$) resulted from the ExpHi clusters being produced more accurately than ExpLo clusters, while there was no effect of Experimental Frequency for the HiEng clusters (F (1,15) = 0.018, $p = .896$, $\eta_p^2 = .001$). Figure 4.2 provides a graph of the accuracy results.

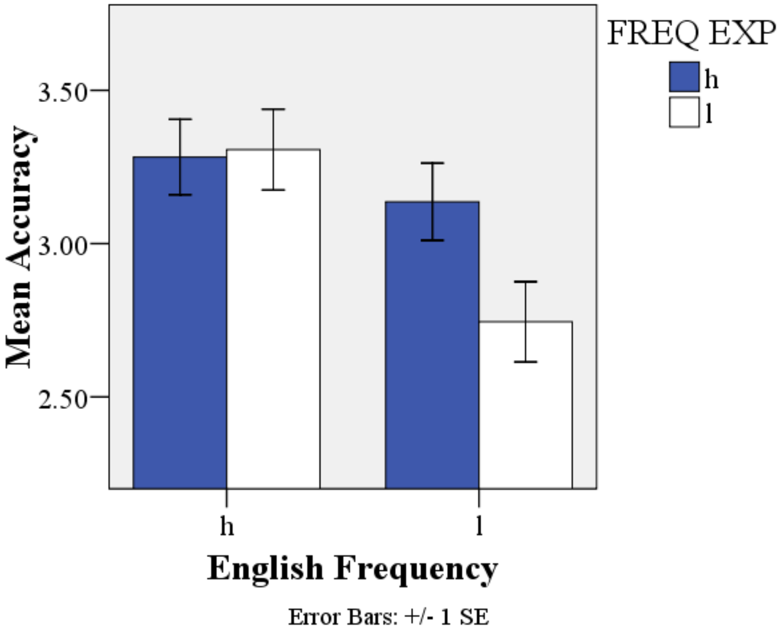


Figure 4.2: Bar graphs of the accuracy results from Experiment 2A. On the ordinate is the mean accuracy for the medial consonant clusters, with a score of ‘4’ being accurate production of both consonants. The bars are shaded according to Experimental Frequency and grouped by English Frequency.

For the analysis of production latencies, there were no significant effects (English Frequency: F (1,15) = .129, p = .725, η_p^2 = .009; Experimental Frequency: F (1,15) = 0.064, p = .804, η_p^2 = .004; English Frequency × Experimental Frequency: F (1,15) = .039, p= .846, η_p^2 = .003). Figure 4.3 provides a graph of the results. Given the interaction present in the accuracy analysis, these results suggest that the production latencies measurement is not as sensitive to learning effects as the accuracy measurement is. This point will be taken up again in the discussion in Chapter 6.

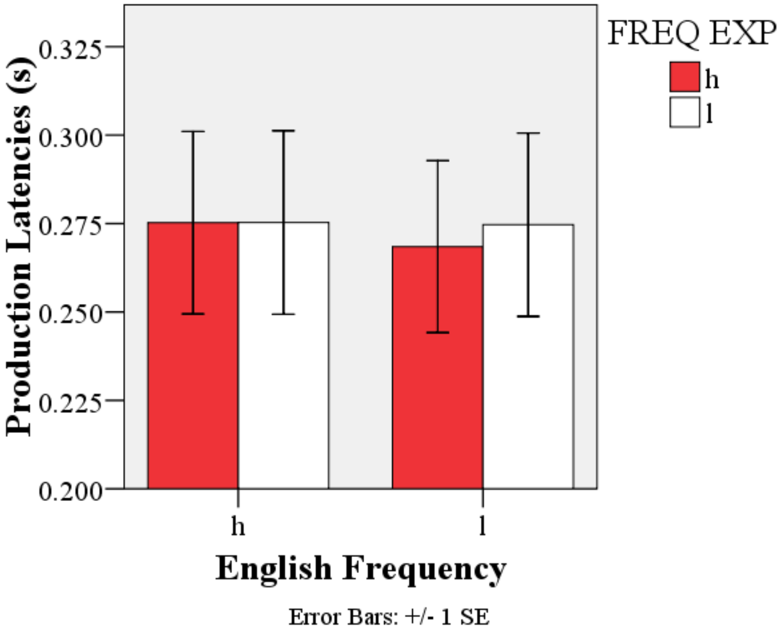


Figure 4.3: Bar graphs of the production latency results from Experiment 2A. On the ordinate is the time from the end of the target word to the onset of the child's production in seconds. The bars are shaded according to Experimental Frequency and clustered by English Frequency.

Ceiling Effects

Histograms of the accuracy scores are given in Figure 4.4. To determine whether there was a ceiling effect present in the accuracy measurement, average accuracy scores for each subject for the HiEng and LoEng clusters were analyzed using a Kolmogorov-Smirnov test for normality, the results of which was combined with visual inspection of the histograms to determine whether a ceiling effect was likely to be present in the data. The HiEng cluster data differed significantly from a normal distribution (D (df = 16) = .243, $p < .05$), but the LoEng clusters did not (D (df = 16) = .119, $p \geq .200$). It is possible, given these results, that effects of the Experimental Frequency manipulation were hidden for the HiEng clusters, but there is no evidence that the LoEng data were limited in their ability to spread across the possible values.

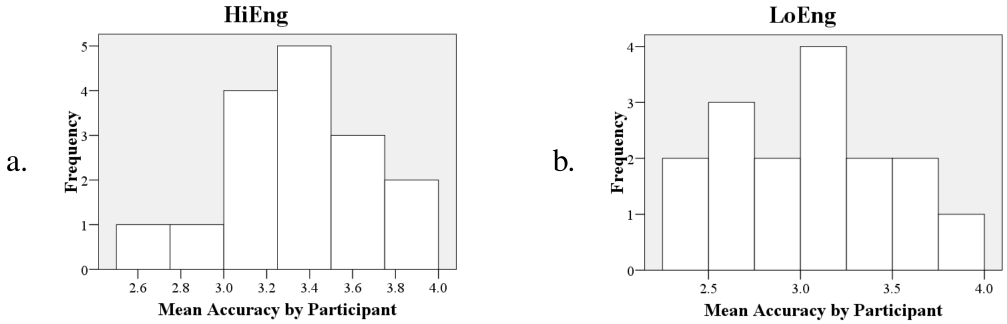


Figure 4.4: Histograms of the accuracy measurements for the HiEng and LoEng conditions in Experiment 2A. Accuracy scores for the HiEng clusters are given in 4.4a and scores for the LoEng clusters are given in 4.4b.

4.1.6 Discussion

Consistent with Experiment 1 and as expected based on previous research on the influence of phonotactic probability on speech accuracy, children were more accurate when producing consonant clusters with high phonotactic probabilities (HiEng > LoEng). However, unlike in Experiment 1, children's productions in Experiment 2A were also influenced by Experimental Frequency. Specifically, children were more accurate at producing the LoEng clusters when they heard three words containing that cluster during the familiarization (LoEng+ExpHi > LoEng+ExpLo). This finding suggests that children were able to learn those phonotactic sequences during the perceptual familiarization and generalize them to the production of new words. Thus, the LoEng clusters provide evidence that type frequency does generalize.

Compare the present results to the results of Experiment 1. In that experiment, the manipulation of token variability did not influence children's productions of the target phonotactic sequences in new words, leading to the conclusion that perceptual learning may not be relevant to learning phonotactic probabilities. The present results show otherwise; they indicate that perceptual learning can allow for generalization to speech and suggest that type frequency is an important component to phonotactic learning.

To see why type frequency was so important, consider the fact that token variability was present in both experiments. Children heard the ExpHi items in Experiment 1 spoken by 10 talkers, and they heard every word in Experiment 2A spoken by four talkers. However, while token variability on its own was ineffectual, the addition to Experiment 2A of several unique words, or types, containing a cluster significantly modulated children's production accuracy. Therefore, it appears that perceptual learning is dependent upon type frequency. This allows us to explain the null effect of Experimental Frequency from Experiment 1 while maintaining the relevance of perceptual learning: Experimental Frequency, and hence perceptual learning, did not have an effect in Experiment 1 because it did not involve multiple types.

Claims about the phonological value of type frequency have been made in a number of phonological domains, including in morphological processing (Albright & Hayes, 2003; Bybee, 1995, 2001) and in phonotactic learning (Albright, 2007; Pierrehumbert, 2003b). The present results tighten these models by adding specificity to the mechanism leading to phonotactic effects. That is, perceptual learning is a necessary component to the development of phonotactic frequency in speech production.

Moving to the discussion of all three models of phonotactic learning, every model made the same prediction, that generalization should be possible when both type frequency and token variability are present in the same experiment. Broadly speaking, all three models were validated by the results. An exemplar model such as the Generalized Neighborhood Model (Bailey & Hahn, 2001) can handle these results adequately by positing connections between the tokens of different words, which are created during the familiarization phase and are subsequently activated by the related test word. As we saw in Experiment 1, however, connections between word-tokens are not sufficient for generalization to take place, so from the perspective of an exemplar-based model, it is unclear why children were able to generalize in Experiment 2A but not in Experiment 1. Are different types of connections made across words compared to within words? This is an answer that is consistent with exemplar models in general, but it is also counterintuitive and not a prediction made by any model that I have found in the literature. Furthermore, only the Albright (2007) and Pierrehumbert (2003b) models make the strong prediction that type frequency is *necessary* for generalization. The results certainly support this claim, and so we may take the results as additional evidence in favor of type-based models of phonological learning.

The findings provide new insights about why phonotactic frequency effects are found in young children's productions. Studies by Beckman and Edwards (1999), Edwards et al. (2004), Munson (2001), Storkel and Rogers (2000), Storkel (2004), and Zamuner et al. (2004) have shown that children are more accurate at producing high frequency phonotactic sequences compared to low frequency sequences, but those authors did not focus on the mechanism by which productive knowledge of phonotactics could be learned. Gerken et al. (2006) subsequently showed that children's productions of a particular phonotactic sequence were modulated by hearing that sequence spoken by multiple talkers, providing evidence that productive phonological knowledge is gained through perceptual learning. In that study, however, children were familiarized and tested with the same words, so it was unclear whether children learned something about phonotactics or simply about whole words. If children learn about sublexical word structure from token variability, we would expect them to be able to genralize that knowledge to new words. The results from Experiment 1 compared with Experiment 2A, however, suggest that the children in the 10 Talkers study and Experiment 1 were learning representations specific to the familiarization words that could not be applied elsewhere.

In contrast, the results from Experiment 2A suggest that perceptual learning *can* lead to productive phonological knowledge. Because children were asked to produce novel words, they were forced to rely on knowledge from the familiarization that was specific to the the medial clusters and broadly true of all the related familiarization words. These results can be interpreted as a natural consequence of the kinds of perceptual learning seen in the infant perception literature. For example, Jusczyk, Friederici, et al. (1993) and Jusczyk et al. (1994) both showed that infants are sensitive to phonotactic patterns before their first birthday. Work done by Boysson-Bardies and colleagues (Boysson-Bardies et al., 1984, Boysson-Bardies et al., 1989, show a clear influence of ambient language patterns in babbling and a child's first words. The present results suggest that perceptual learning is not isolated to infants, but continues to be relevant as children begin to speak. Given that perceptual learning can be shown to affect speech production, it is highly likely that perceptual learning is an important component to speech and motor planning. The present study shows how perceptual learning affects production accuracy, but the 10 Talkers study found that perceptual learning also affected production speed. Although no such effect was found here, further research may help to determine where and when perceptual learning can be expected to affect production speed. Additionally, we may find that perceptual learning is relevant to speech flexibility (Munson, 2001) and speech variability (Goffman, Gerken, & Lucchesi, 2007).

Interpreting the Interaction

Returning now to the lack of an effect of Experimental Frequency in the HiEng clusters, there are several factors that may have contributed to the interaction in the accuracy analysis. One possibility is a ceiling effect. Given the demands of the task and the efficacy of perceptual learning, there is some chance that children's performance on HiEng clusters could be not be raised much beyond about 3.3. The results of the Kolmogorov-Smirnov test for normality support the claim that a ceiling effect limited the spread of accuracy scores for the HiEng clusters. In consequence, the significant difference between Experimental Frequency conditions for the LoEng clusters would have resulted because average performance on these words was significantly lower. This claim must be made relative to the task demands, however, because average scores in Experiment 1 were actually higher compared to those obtained here.

Another explanation for the interaction concerns the nature of the HiEng word set. Given that the word /zaspən/ was dropped (reducing the item set from eight to seven), it is possible that the effect of Experimental Frequency was hidden by limited data and low power.

The results are also consistent with an interaction of Experimental Frequency and syllable structure. The fact that Experimental Frequency only affected the LoEng clusters significantly suggests that perceptual learning affects licit consonant clusters one way (i.e., clusters that can be parsed as a single syllabic entity, such as an onset or a coda) and illicit clusters another way. In other words, perceptual learning may be more relevant to clusters which cannot be parsed as either onsets or codas.

The most satisfactory explanation for the interaction, however, is based on the ceiling effect. There is no obvious reason why differences in syllable structure would predict an effect for the LoEng clusters and not for the HiEng clusters. It seems equally likely, for example, that the phonological status of the HiEng clusters would make them *more* susceptible to perceptual learning than LoEng clusters. The histograms in Figure 4.4, however, provide some evidence that children were producing the HiEng clusters at or near ceiling.

Regardless of why an interaction was found, what is most relevant is that some degree of generalization occurred. Recall that the primary purpose of the English Frequency factor was to corroborate previous findings on phonotactic probability effects in child speech, and to provide a context within which the Experimental Frequency manipulation can be interpreted. Whether perceptual learning in this experiment was obscured by a ceiling effect or was modulated by ambient language factors, the results nevertheless support some degree of phonotactic learning from a set of relevant words in the ambient language. This fact may be extended to explain why the English Frequency effects exist in the first place, namely, because of perceptual learning. This learning may be limited, as a syllble-structure based interpretation of the interaction would suggest, or the perceptual learning effects may have been limited, as the ceiling effect interpretation would suggest. Either way, it is clear that perceptual learning contributes to phonotactic learning. Furthermore, the effect of Experimental Frequency is being interpreted by comparison with the results from Experiment 1, in which no effect of perceptual learning was found. Although a ceiling effect may also have occurred with respect to the HiEng data in Experiment 1, there is no evidence that a ceiling effect influenced the LoEng clusters, for which the learning effects were seen here. We may therefore conclude that Experiment 2A and the manipulation of type frequency provide evidence for generalization, whereas the manipulation of token variability in Experiment 1 does not.

Other concerns about the results can be raised, however. As mentioned in the materials section, the ExpHi and ExpLo familiarization words were divided such that that ExpLo words only began with labials and the ExpHi words rarely started with labials. Although this design helped narrow the possible generalizations present during the familiarization to just the medial clusters, it also resulted in a confound. It is unlikely that the results were greatly influenced by the initial consonant, but it remains a possibility.

Concerns can also be raised for the word /simpən/. Accuracy results for this word were the lower than for all other words, as can be seen in Figure 4.5. These low scores may be related to a neutralization of nasal place of articulation contrasts before high vowels[4]. The primary concern with respect to this item, then, is that it contributed unwanted noise to the results with respect to both the English Frequency and Experimental Frequency factors.

[4] The formant transitions into a nasal consonant are important cues for the nasal's place of articulation (Stevens, 1998). In high vowels, these transitions are severely restricted, particularly the F2 transition, which differentiates bilabials and velars, making contrasts between the /m/ and the /ŋ/ difficult to perceive (B. Story, personal communication).

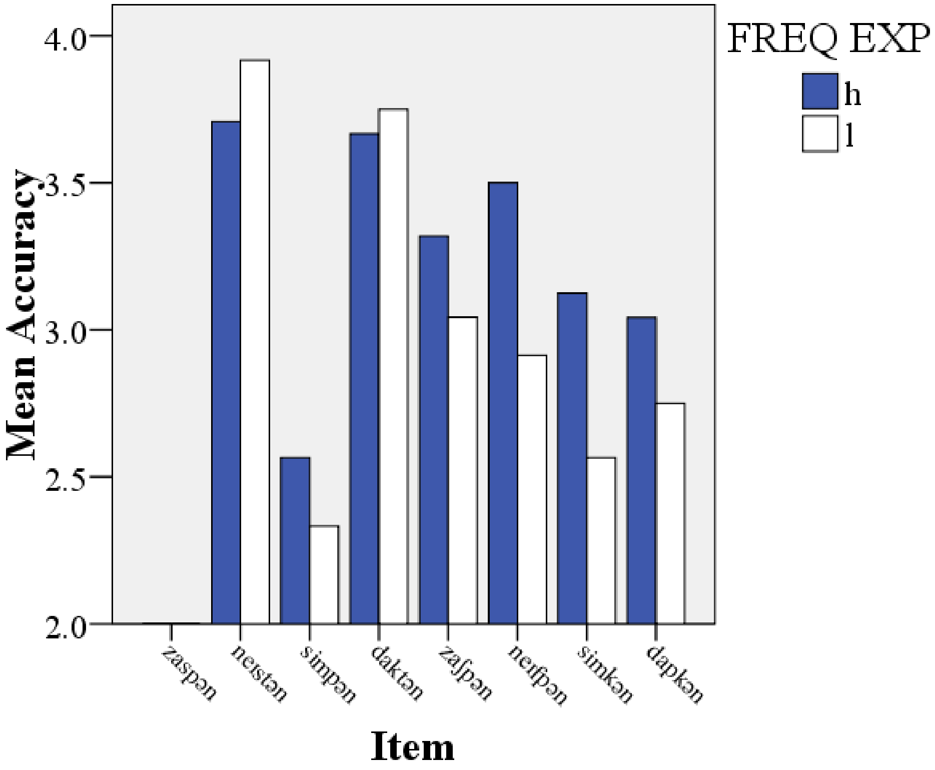


Figure 4.5: Accuracy results for each test word from Experiment 2A, broken up by Experimental Frequency. The item /zaspən/ was removed from the analysis, so no results for this item are shown.

A final concern relates to the phonotactic probabilities of the three word sets (the ExpLo, ExpHi and test sets). Although phone and biphone sums were shown to be relatively even across the three sets, individual phones and biphones were not controlled. That is, the initial consonants, the initial CVs, and all the other phonotactic sequences besides the medial clusters, may not have been balanced across the word sets. Therefore, the results may reflect some oddity related to individual phones or biphones that were not matched across the three sets, rather than a generalization about the medial clusters. Any of these possibilities may have led to the absence of an effect of Experimental Frequency in the HiEng condition or otherwise influenced the results. These concerns are addressed in the next experiment, 2B, discussed in the next chapter.

CHAPTER 5

EXPERIMENT 2B: PHONOTACTIC GENERALIZATION FROM MULTIPLE WORD-TYPES AND MULTIPLE TALKER-TOKENS

The results of Experiment 2A suggest that children are able to generalize knowledge of a phonotactic sequence based on perceptual learning, at least for low probability phonotactic sequences. Several possible concerns were raised in the discussion, however. The purpose of Experiment 2B was to address these concerns and, at the same time, confirm that perceptual learning effects can be found in speech production.

As in Experiment 2A, children were familiarized with word-medial consonant clusters in either one (ExpLo) or three (ExpHi) nonsense words. Each word was produced by four different talkers. Following the familiarization, children were presented with new words containing the same medial clusters and were asked to produce those words. In Experiment 2B, place and manner of articulation were distributed across the ExpHi and ExpLo word sets by making the ExpLo set a subset of the ExpHi set. Furthermore, the three word sets (ExpLo, ExpHi, and test word sets) were all matched for individual phone and biphone scores, as well as for phone and biphone sums for whole words. Therefore, the words in Experiment 2B were highly constrained.

The predictions made by the exemplar-based model, Pierrehumbert's model, and Albright's model all remain the same: all three models predict that generalization of a phonotactic sequence should be possible when participants are presented with multiple word-types and multiple word-tokens containing the same phonotactic sequence. These predictions are repeated in Table 5.1 below.

Table 5.1: Predictions made by the three models, Bailey and Hahn's exemplar model, Pierrehumbert's hybrid model, and Albright's MGL, for the second word-types + word-tokens experiment. *Gen* is given in cells where a model (row) predicts a significant effect of the Experimental Frequency manipulation for a given experiment (column). *No Gen* refers to a predicted null effect. A prediction that is unique to a particular model is boxed and in bold.

	EXP 1 Tokens	EXP 2A Types + Tokens	EXP 2B Types + Tokens	EXP 3 Types
EXEMPLAR MODEL Bailey and Hahn (2001)	**Gen**	Gen	Gen	Gen
HYBRID MODEL Pierrehumbert (2003b)	No Gen	Gen	Gen	**No Gen**
MINIMAL GENERALIZATION LEARNER Albright (2007)	No Gen	Gen	Gen	Gen

5.1 Method

5.1.1 Participants

Twenty-six children participated in the study. The participants were between the ages of 4;0 and 4;3 (mean = 4;1.3) and were recruited using the same database of local birth announcements and adoptions in Tucson that was used in Experiments 1 and 2A. All were native English speakers, and had minimal exposure to other languages. Exposure to a second language was always less than 10 hours per week. The average was 0.94 hours/week, the mode was 0 hours/week. Nine children had exposure to Spanish, two to ASL, one to German, one to French, and one to Russian. Five children were reported by their mothers to have a personal or family history of speech, hearing, language, or cognitive deficits, and five others did not complete the experiment or did not follow instructions. These children were removed from the analysis. Results from the remaining 16 participants, 9 girls and 7 boys, are reported below.

5.1.2 Materials

As in Experiment 2A, the target phonotactic sequences were again the medial clusters of CVCCVC nonsense words. The word were divided into three sets, two familiarization sets and one test set, and varied according to two factors: English Frequency and Experimental Frequency. For the English Frequency factor, one group of clusters had high phonotactic probabilities in English (HiEng) and the other group had low or null probabilities (LoEng). Experimental Frequency was a manipulation of the number of times that a cluster was heard during the familiarization. The ExpHi familiarization set contained three words corresponding to each cluster, the ExpLo familiarization set contained only one word per cluster. The test word set also contained one word per cluster.

Familiarization Words

The familiarization word lists are given in Table 5.2. The Experiment 2B word list contains several changes to the word list used in Experiment 2A, the purpose of which was to provide greater control and limit word differences to the important manipulations, English Frequency and Experimental Frequency.

In the previous experiment, the majority of ExpHi words started with non-labial consonants that never appeared at the beginning of ExpLo words (for example, /dompət/, /lɛktəf/, /kɛspəs/, and /sʌstəp/). Likewise, the ExpLo words had only labial onsets (/foʃpəm/, /mæfpəm/, /fospəm/, and /mæstəm/). The purpose of this setup was to avoid unwanted generalizations that might have existed across ExpLo and ExpHi words, such as a generalization about the initial consonant. Unfortunately, using separate word-initial consonants for the ExpHi and ExpLo sets also resulted in a confound. In the present experiment, the same words were used in both lists, so the same word-initial consonants appeared in both the ExpLo and ExpHi word sets. A second important change to the present word list was that individual phone and biphone probabilities were compared across the three lists, and words were chosen such that these probabilities were equated. The purpose of these changes was to eliminate spurious variability across word sets and ensure that the English Frequency and Experimental Frequency manipulations could be

Table 5.2: Familiarization words used in Experiment 2B. The distribution of clusters across the two lists and the two blocks (*block* = a familiarization phase followed by a test phase, two blocks per session) are arranged identically to Experiment 2A. Items are written in IPA.

	ENG FREQ	LIST 1		LIST 2		EXP FREQ
Block One	**HiEng**	**mp**	**kt**	**sp**	**st**	*ExpHi*
		dɪmpət	lɛktəf	kɛspəs	mæstəm	
		nʌmpəs	saʊktəs	tuspən	baɪstəm	
		gumpən	biktəm	fospəm	gɪstək	
	LoEng	**ʃp**	**fp**	**mk**	**pk**	*ExpLo*
		foʃpəm	mæfpəm	nʌmkəs	saʊpkəs	
Block Two	**HiEng**	**sp**	**st**	**mp**	**kt**	*ExpLo*
		fospəm	mæstəm	nʌmpəs	saʊktəs	
	LoEng	**mk**	**pk**	**ʃp**	**fp**	*ExpHi*
		dimkət	lɛpkəf	kɛʃpəs	mæfpəm	
		nʌmkəs	saʊpkəs	tuʃpən	baɪfpəm	
		gumkən	bipkəm	foʃpəm	gɪfpək	

interpreted as clearly as possible.

English Frequency

The familiarization and test words in Experiment 2B were equated for the phone and biphone sums of each word, as well as for individual phones and biphones. The Phonotactic Probability Calculator (Vitevitch & Luce, 2004) was again used to find the phone and biphone scores for each word Set (ExpLo, ExpHi, and Test). Individual phone and biphone probabilities for the familiarization word /mæfpəm/ are given in Table 5.3. Phone1-Phone6 correspond to the phonotactic probability scores for the six component phonemes; Biphone1-Biphone5 correspond to the scores for the five component biphones. Phone Sum and Biphone Sum are the sums of these scores over the entire word.

The phone and biphone scores and the phone and biphone sums were then balanced across lists so that the scores for the words in one list were comparable to the scores of the related word(s) in another list. For example, the individual phones and biphones, as

Table 5.3: Individual phone and biphone scores, as well the the phone and biphone sums, for the familiarization word /mæfpəm/.

Phone1 **m**	Phone2 **æ**	Phone3 **f**	Phone4 **p**	Phone5 **ə**	Phone6 **m**	Phone Sum **mæfpəm**
.0572	.0794	.0197	.0362	.0816	.0355	.3096

Biphone1 **mæ**	Biphone2 **æf**	Biphone3 **fp**	Biphone4 **pə**	Biphone5 **əm**		Biphone Sum **mæfpəm**
.0101	.0013	.0000	.0042	.0117		.0272

well as the phone and biphone sums, of the ExpLo word /mæfpəm/ were equated with the scores for the ExpHi words /baɪfpəm/, /gɪfpək/, and /mæfpəm/, and for the test word /neɪfpən/. The raw probabilities for each set, organized by the relevant scores and clusters, are provided in Appendix C.

To compare the three word sets, scores for the six component phones (Phone1-Phone6), the five component biphones (Biphone1-Biphone5), the sum of the phone scores for each word (Phone Sum) and the sum of the biphone scores (Biphone Sum) were each entered as variables into an ANOVA with Set and English Frequency as between-item factors to determine whether the word sets were balanced. No significant effects of the Set variable were found (all F s < 1) for any of the phonotactic probability scores. English Frequency was significant for Phone3 (F (1,18) = 27.208, $p < .001$), Phone4(F (1,18) = 6.995, $p < .05$) and the Phone Sum (F (1,18) = 16.032, $p < .01$); there were also significant effects of Biphone3 (F (1,18) = 19.833, $p < .001$), Biphone4 (F (1,18) = 17.236, $p < .01$), and the Biphone Sums (F (1,18) = 10.836, $p < .01$), as well as a trend towards significance for Biphone 2 (F (1,18) = 4.115, p = .058). There was no significant Set × English Frequency interaction for any of the variables (all Fs < 1). The significant effect of English Frequency is consistent with the manipulation of the medial consonant cluster, and in all cases, the HiEng probabilities were greater than the LoEng probabilities. Specific results from the analyses of Set and English Frequency are included in Appendix C. The results of these tests confirm that the word sets were balanced with respect to the phone and biphone sums as well as for individual phones and biphones.

Syllable structure In Experiment 2A there was an English Frequency × Experimental Frequency interaction that resulted from LoEng clusters being produced more accurately when they were associated with three words compared to just one word. The interaction suggests that syllable structure may have been influencing the results, although the interpretation of this interaction is mitigated by several concerns, discussed in detail in the previous chapter, and the likelihood that facilitation was not seen for the HiEng clusters because children were producing those words at ceiling. If syllable structure is influencing the results, we would expect similar behavior for words with similar syllable structure, as well as for the same clusters across experiments. These possibilities are discussed in greater detail in the discussion, as well as in Chapter 7.

Experimental Frequency

Experimental Frequency was a manipulation of the number of times that a cluster was heard during the familiarization. During the familiarization phase of the experiment, ExpHi clusters were heard in three words, ExpLo clusters were only heard in a single word. Clusters were assigned to ExpHi in one word list and to ExpLo in a second list so that half of the participants heard a cluster as ExpHi and the other half heard it as ExpLo. The lists were also created with other issues of balance in mind, including the removal of potentially distracting generalizations about word-initial CVs, the number of talkers heard by participants in this experiment and in Experiment 3, and the possibility that overly similar words might create online neighborhood effects during the experiment. The steps taken to address these concerns are described next.

Word-initial consonants and vowels The division of ExpHi and ExpLo sets across lists is given in Table 5.2 and is identical to the division in Experiment 2A (cf. Table 4.2 for a comparison). The purpose of the experiment was to have children learn generalizations about the medial consonant clusters, so it was important to avoid unwanted generalizations, such as might be true of word-initial consonants. In Experiment 2A spurious generalizations were eliminated by assigning different word onsets to the ExpHi and ExpLo sets, but this solution to unwanted generalizations resulted in a confound of initial consonant and Experimental Frequency.

As mentioned previously, the same consonants were used in both the ExpHi and ExpLo sets in Experiment 2B. This can be seen by comparing the ExpLo words in Block One of List 1 (/foʃpəm/ and /mæfpəm/) with the ExpHi words in Block Two of List 2 (/kɛʃpəs/, /tuʃpən/, /foʃpəm/, and /mæstəm/, /baɪfpəm/, and /ɡɪfpək/). The ExpLo words in Block One are reused as one of the three ExpHi words in Block Two of List 2. In other words, the ExpLo words were a subset of the ExpHi words, so the same places and manners of articulation appeared in all Experimental Frequency conditions. This eliminated the initial-consonant confound. Table 5.2 also shows how unwanted generalizations were again avoided: each word in a given block of the familiarization phase started with a unique consonant. For example, the word-initial consonants in Block One of List 1 were /d l n s ɡ b f m/, all unique word onsets. Additionally, the first vowel of each familiarization word was unique. In sum, the Experimental 2B materials were designed so that there was no place of articulation confound and the medial consonant clusters represented the only generalizations possible within an experimental block.

Talkers As in Experiments 1 and 2A, a different talker was used for each of the four tokens of each word. Eight women were recorded in total; all spoke a Western US dialect of American English. The recording setup used in Experiments 1 and 2A was used here, and care was again taken to ensure that each token used in the familiarization had clear phonetic cues for the medial consonant cluster and had trochaic stress. The tokens chosen for use in the experiment were equated for root-mean-square amplitude. Because eight total talkers were used, each talker was associated with four of the eight familiarization words in a block. This allowed for the total number of talkers to be balanced across Experiments 2B and 3. See Section 6.1.2 of Chapter 6 for further details.

Neighborhood density As discussed in previous chapters, the correlation between phonotactic probability and neighborhood density necessitates the control of neighborhood density across word sets and the English Frequency conditions.

Neighborhoods for the word sets were determined using the one-phoneme edit distance metric (Luce, 1986) in the Washington University Neighborhood Database (http://128.252.27.56/Neighborhood/Home.asp). None of the familiarization or test words had lexical neighbors with the exception of /nʌmpəs/, which has the neighbors *compass* and *rumpus*. These neighbors are relatively low frequency (13 and 1 occurrences per million, respectively). Sparse neighborhoods and low-frequency neighbors have minimal effects on online word recognition (Luce & Pisoni, 1998), so they were unlikely to influence or skew perception of /nʌmpəs/, particularly for four-year-old children whose vocabularies are not yet fully developed. It is also unlikely that an online neighborhood might have been created (Hollich et al., 2002), particularly because of the unique CV onset assigned to each familiarization word in a given block.

Familiarization Lists

The two word lists were again subdivided (not shown in Table 5.2) so that the clusters /kt/, /mp/, /ʃp/, and /fp/ were heard in Block 1 for one list and in Block 2 for the other list, making a total of four lists that were used in the experiment (1A, 1B, 2A, and 2B). All four lists are given in Appendix A.

Test Words

The test words used in Experiment 2B are given in Table 5.4. The test words used in Experiments 1 and 2A were updated so that phone and biphone probabilities were similar to the phone and biphone probabilities of the familiarization sets. The results of the ANOVA validating this claim are given in the description of the English Frequency factor above. The test words were recorded in the same manner described above for the familiarization words using a novel talker (female, native speaker of American English with a Western US accent). This talker was also asked to produce the words with trochaic stress and in a manner such that important phonetic cues to each consonant in the cluster (release of the first stop in the cluster, aspiration on the second stop) were clearly audible in the test word token. Representative tokens were chosen from the talkers recordings and were normalized for root-mean-square amplitudes.

Table 5.4: Test words created for Experiment 2B. These words are matched with the familiarization word sets for individual phone and biphone scores and for phone and biphone sums. Items are written in IPA.

High English	**Low English**
sp /daspək/	**ʃp** /daʃpək/
st /neɪstən/	**fp** /neɪfpən/
mp /sæmpəf/	**mk** /sæmkəf/
kt /tuktən/	**pk** /tupkən/

The test words also differed with respect to their initial CVs compared to the words in the ExpHi and ExpLo word sets, which eliminated the possibility of unwanted generalizations present across familiarization and test words.

5.1.3 Procedure

The procedure for Experiment 2B was identical to that used for Experiment 2A.

5.1.4 Analysis

The same dependent measures used in Experiment 1 were collected here: accuracy on the two medial consonants and production latencies. Accuracy scores were out of four possible points for the cluster. Transcriptions were made by two data coders working independently. Inconsistencies between the transcribers were either retranscribed by agreement or were removed from the analysis. Less than 2% of the data was removed due to a lack of agreement, and never more than three items were removed per participant.

Production latencies were determined by first finding the length of the target word in the sound file, then finding the time from the end of the target to the beginning of the child's own production. Measurements were made using Praat software (www.praat.org) and entered into a spreadsheet for analysis. Measurements were made by the two coders independently, and agreement between them was high: 83.3% of their measurements were within 50 ms of each other, and 92.7% were within 100 ms. Approximately 6% of the data was removed because children attempted the production twice.

As in Experiments 1 and 2A, the first repetition of each word was removed from the production latency analysis due to the high level of variance in those productions. Figure 5.1 gives the average accuracy and production latency scores for each repetition collapsed across subjects and words. The production latencies for the first repetition were again significantly longer than any of the other repetitions. Accuracy results across the four repetitions were relatively consistent, so all of these data were included in the analysis.

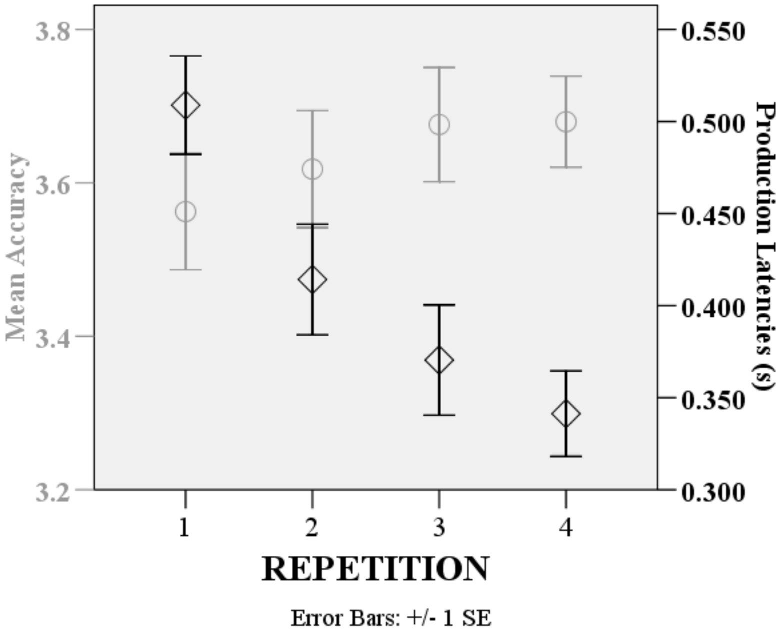


Figure 5.1: Average accuracy and production latency scores for each of the four repetitions collected in Experiment 2B. The accuracy averages are represented by open circles and are shaded gray. The black diamonds are the averages for the production latencies. The scale for the accuracy averages is on the left, the scale for the production latency averages is on the right.

5.1.5 Results

Table 5.5 gives mean accuracy and production latency scores and standard deviations for each of the four conditions in Experiment 2B.

Table 5.5: Mean accuracy (out of 4) and production latency (in seconds) scores for Experiment 2B. Standard deviations are given in parentheses.

Accuracy				
	HiEng ExpHi	*HiEng ExpLo*	*LoEng ExpHi*	*LoEng ExpLo*
Mean	3.715	3.885	3.598	3.314
(Std Dev)	(.393)	(.187)	(.432)	(.388)

Production Latencies				
	HiEng ExpHi	*HiEng ExpLo*	*LoEng ExpHi*	*LoEng ExpLo*
Mean	.384	.432	.392	.426
(Std Dev)	(.108)	(.124)	(.126)	(.150)

For the accuracy analysis, there was a significant effect of English Frequency (F (1,15) = 27.527, $p < .001$, $\eta_p^2 = .647$) resulting from greater accuracy scores for HiEng words compared to LoEng words. This effect held for the individual data from 14 of the 16 participants, indicating that the results were generally true of the population. There was no effect of Experimental Frequency (F (1,15) = 0.363, $p = 0.556$, $\eta_p^2 = .024$), but there was a significant English Frequency × Experimental Frequency interaction (F (1,15) = 11.35, $p < .01$, $\eta_p^2 = .431$). Comparing the simple effects, a significant effect of Experimental Frequency for the LoEng clusters (F (1,15) = 4.24, $p < .05$, $\eta_p^2 = .247$) resulted from the ExpHi clusters being produced more accurately than ExpLo clusters. This effect held for the individual data from 12 of the 16 participants. There was not a significant effect of Experimental Frequency for the HiEng clusters (F (1,15) = 2.701, $p = .121$, $\eta_p^2 = .153$). Figure 5.2 presents a graph of these results.

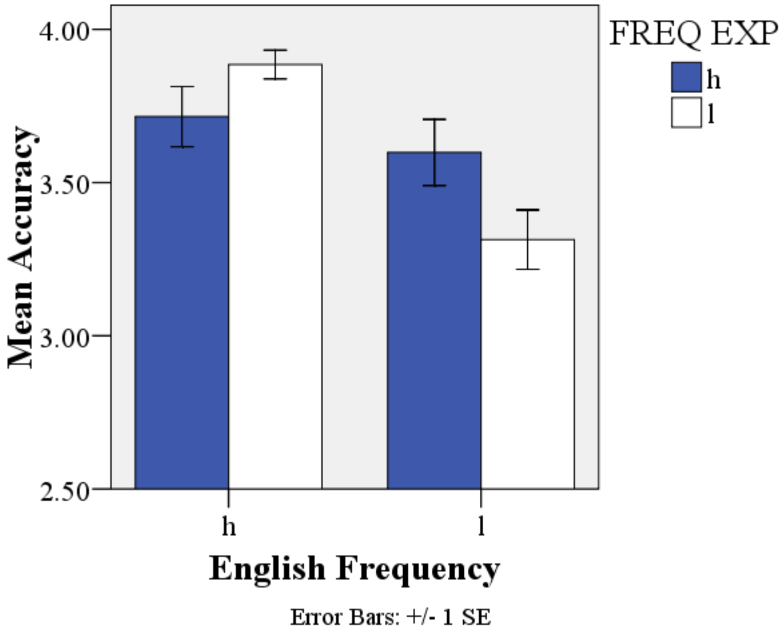


Figure 5.2: Bar graphs of the accuracy results from Experiment 2A. On the ordinate is the mean accuracy for the medial consonant clusters, with a score of '4' being accurate production of both consonants. The bars are shaded according to Experimental Frequency and grouped by English Frequency.

For the analysis of production latencies, there were no effects of English Frequency (F (1,15) = 0.005, p = .945, η_p^2 = .000) or Experimental Frequency (F (1,15) = 2.846, p = 0.112, η_p^2 = .159), and no interaction (F (1,15) = 0.093, p= .765, η_p^2 = .006). Consistent with Experiment 2A and findings reported by Munson et al. (2005), it does not appear that the production latency measurement is particularly sensitive to either the English Frequency or Experimental Frequency manipulations.

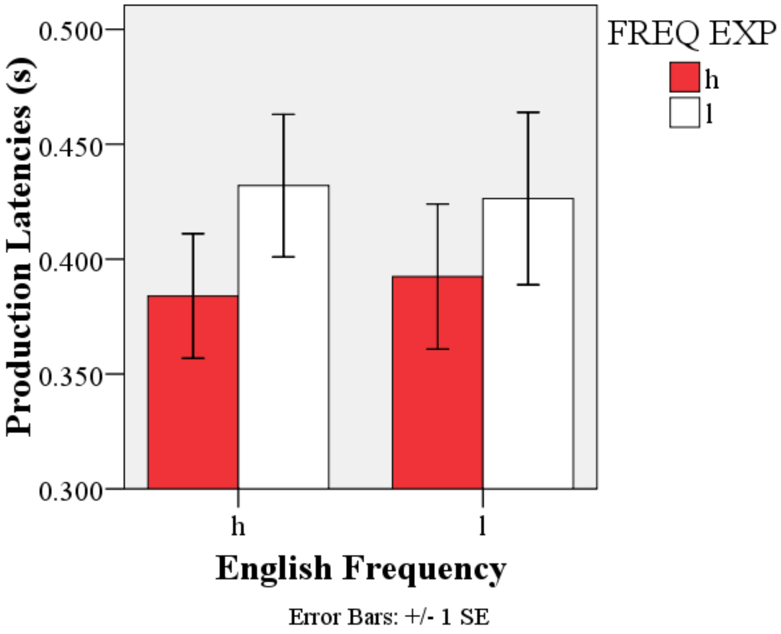


Figure 5.3: Bar graphs of the production latency results from Experiment 2B. On the ordinate is the time from the end of the target word to the onset of the child’s production in seconds. The bars are shaded according to Experimental Frequency and clustered by English Frequency.

Ceiling Effects

Histograms of the accuracy scores are given in Figure 5.4. To determine whether there was a ceiling effect present in the accuracy measurement, average accuracy scores for each subject for the HiEng and LoEng clusters were analyzed using a Kolmogorov-Smirnov test for normality. The HiEng cluster data differed significantly from a normal distribution (D (df = 16) = .246, $p < .05$) resulting from a reduced spread above the mean, but no significant difference from a normal distribution was found for the LoEng clusters (D (df = 16) = .135, $p \geq .200$). It is possible, given these results, that effects of the Experimental Frequency manipulation were hidden for the HiEng clusters, but there is no evidence that the LoEng data was limited in its ability to spread across the possible values.

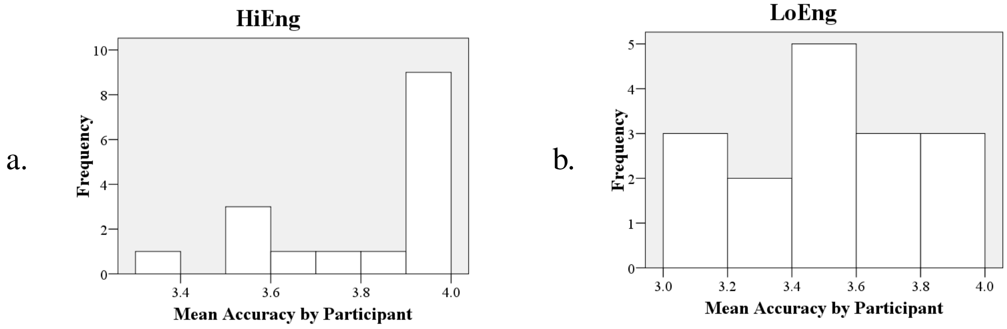


Figure 5.4: Histograms of the accuracy measurements for the HiEng and LoEng conditions in Experiment 2B. Accuracy scores for the HiEng clusters are given in a. and scores for the LoEng clusters are given in b.

5.1.6 Discussion

The combined results of Experiments 2A and 2B provide strong support for the claim that perceptual learning influences speech production, at least for low probability sequences. Children in these studies were more accurate when producing a LoEng cluster if they first heard three words containing that cluster (the ExpHi condition). The findings are complementary to research on L2 acquisition in which perceptual learning has been found to facilitate the production of non-native contrasts (Bradlow et al., 1997; Gerken et al., 2006; Wang et al., 2003). Contrary to claims that perceptual learning is not relevant to the development of speech for first language learners (Messum, 2007), these results indicate it plays both an influential and facilitative role. Although articulatory practice may also serve some additional, beneficial role in speech development, it is clear that it does not act alone.

The results of Experiments 2A and 2B also support a type-based view of phonotactic learning (Albright, 2007; Bybee, 1995, 2001; Pierrehumbert, 2003a, 2001). In contrast to Experiment 1, in which only token frequency was manipulated and no effect of Experimental Frequency was found, the addition of multiple word-types resulted in significantly improved speech accuracy for LoEng clusters. Furthermore, this improved accuracy was consistent, appearing in two separate experiments that used different sets of words.

Perceptual learning of word-types, or unique lexical items, is a likely mechanism behind the correlation between production accuracy and vocabulary size that was found by Edwards et al. (2004). The present results reinforce the claim that phonotactic representations are derived from the lexicon, but they also reveal how phonotactic generalizations learned over the lexicon can be made in real time. That is, the present results show that flexible, context-independent representations can be built on the fly by the immediate abstraction of knowledge of a phonotactic sequence from words encountered in the ambient language.

The claim that experience with multiple words sharing a phonotactic pattern induces the construction of abstract knowledge echos a similar claim made by Edwards et al.. They compared performance on low- and zero-probability phonotactic sequences in children with both large and small vocabularies, and found that the children with larger vocabularies were more accurate than their peers at producing zero-probability sequences. If knowledge of phonotactics is entirely dependent on unanalyzed lexical exemplars, we expect children in both groups to perform equivalently. The fact that children in the Edwards et al. study with larger vocabularies were more accurate suggests that they were not simply storing encountered words, but they generalized over those forms to create context-independent representations. In these studies, as well, children did not learn phonotactic sequences that were bound to existing forms, but were able to generalize their knowledge to a new word. Furthermore, these representations must have been abstract enough to link different modes of language processing: because perceptual learning affected speech production. Therefore, the results provide additional support for the claim that a phonological grammar of learned, abstract forms exists in the minds of speakers.

Edwards et al. raised the question of why some children had larger vocabularies than others. Children in that study were age matched so that differences between groups could be attributed to vocabulary size. The present results pose one possible answer: children who are better perceptual learners may have larger lexicons and thereby support more abstract phonologies. Unfortunately, no measures of children's vocabularies were made for this study, so this claim cannot be examined here. Future research along these lines should include vocabulary measures for subjects, as well as additional tests of production accuracy, so that the relationship between perceptual learning and vocabulary learning can be further explored.

Interpreting the Interaction

As was discussed in Experiment 2A, the most likely explanation for the lack of facilitation of the Experimental Frequency manipulation for the HiEng clusters is a ceiling effect. Considering the distribution of accuracy scores given in Figure 5.4, it seems very likely that children were too accurate in producing the HiEng clusters to allow for a separation of the ExpHi and ExpLo conditions. Children were also more accurate in Experiment 2B than in 2A, averaging closer to 3.8 for the HiEng clusters (compare to the 3.45 average from 2A), but this is likely a result of using different words. Four-year-olds were recruited in the present study so that the results could be compared with the previous experiments presented here, as well as with the experiments from the 10 Talkers study (Gerken et al., 2006). However, future work of this type would benefit from recruiting a younger group of children, such as 3- or 3.5-year-olds, so that a greater degree of spread can be achieved for the accuracy analysis.

The interaction may also be explained by an interaction of perceptual learning and syllable structure, although, as pointed out in the previous chapter, there are no a priori predictions for why these two factors should interact in a particular way. It is also worth noting that clusters with the same syllable structure do not appear to behave similarly. Figure 5.5 presents a graph of the accuracy scores for each word, collapsed across subjects and broken up by the Experimental Frequency conditions. The trend for ExpHi words to be produced more accurately appears to hold for both ExpHi and ExpLo words, with the exception of /sæmpəf/, which was produced less accurately following familiarization with three /mp/ words. The fact that the interaction was due largely to a single item makes it difficult to interpret the interaction as reflecting a factor such as syllable structure. Nevertheless, phonotactic frequency and syllable structure are inseparable with respect to the present design (cf. Section 3.1.2), so it remains for future research to determine how each of these distinct types of phonological knowledge affect and are affected by perceptual learning.

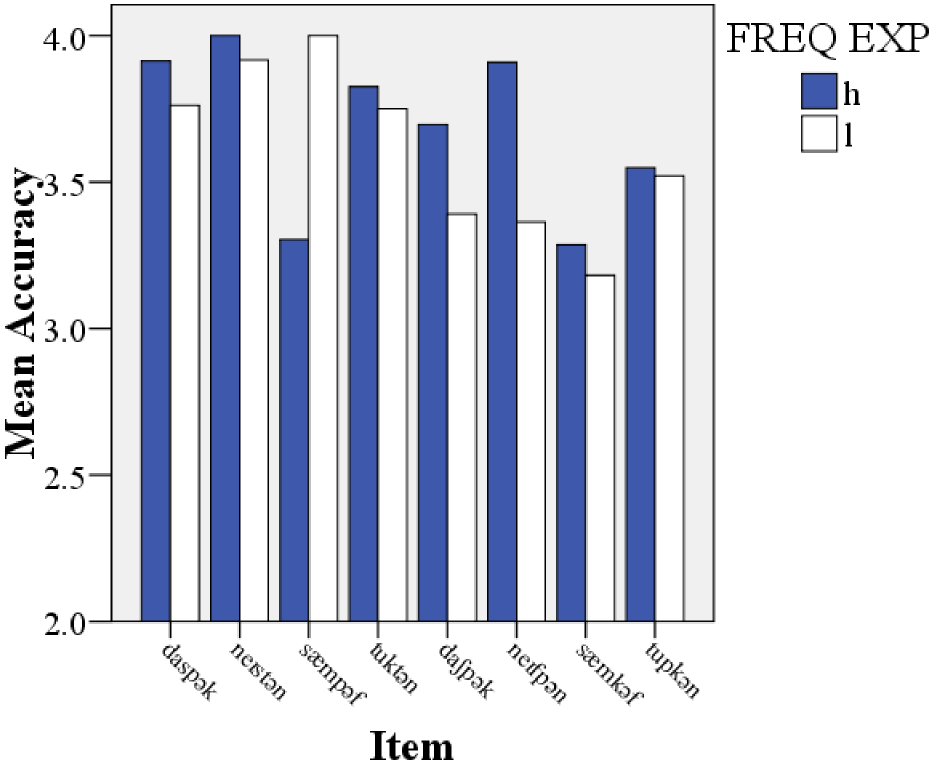


Figure 5.5: Accuracy results for each word broken up by Experimental Frequency.

However future research eventually explains this interaction, the important question relates to the manipulation of Experimental Frequency. The fact that *any* effect of the

Experimental Frequency manipulation was found suggests that perceptual learning is relevant to the development of speech production, and that the effects of hearing multiple words were largely facilitative suggests that the Experimental Frequency manipulation is a smaller, experimentally defined version of English Frequency. As discussed in Chapter 4, the important conclusions come from a comparison of the results of Experimental Frequency across experiments, rather than a comparison of the HiEng and LoEng conditions. It is likely that any manipulation of Experimental Frequency is modulated by what children already know about broader ambient language frequencies, but we may nevertheless conclude that Experimental Frequency does influence speech production.

In the next and final experiment, an additional test of the abstraction hypothesis is made. In particular, I test whether the phonotactic representations children learned were abstract. This is accomplished by manipulating an intermediate level of abstraction that may be necessary for higher level abstractions, namely, the phonetic variability resulting from the use of multiple talker-tokens. This intermediate level consists of abstract word forms. Figure 2.4 (repeated here in Figure 5.6) provides a graphical representation of how abstract word forms are created.

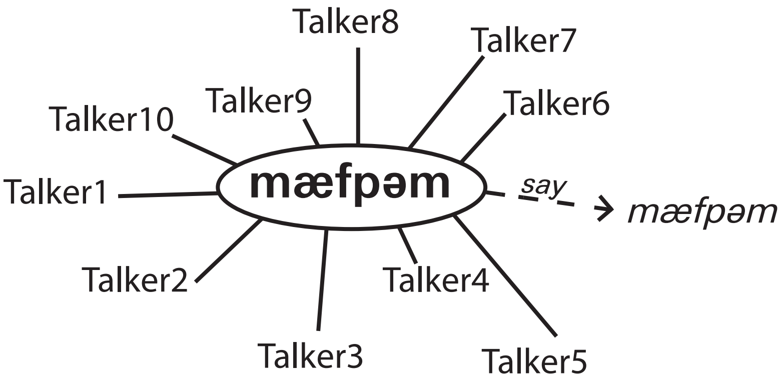


Figure 5.6: Dynamics of Pierrehumbert's (2003) hybrid model for the creation of word-form knowledge.

Exposure to multiple talkers with varying phonetic signatures produce a phonologically coherent unit, the word /mæfpəm/. Listeners are able to abstract away a representation that is common to each encountered token. This process of learning word-forms from phonetic variability is predicted by Pierrehumbert's hybrid model. Furthermore, it is predicted to be a necessary component of phonological learning: phonotactics are learned from the abstract word forms, not from the phonetically variable exemplars. So, if phonetic variability is not present in a perceptual learning experiment, abstract word forms should not be learned, and by transitivity, phonotactics should not learned either.

In Experiment 3, the same familiarization words from Experiment 2B were used, but the talker assignment was changed so that a given word was heard by only a single talker. This removed the level of phonetic variability that resulted in the facilitative effects seen in the 10 Talkers experiment, and is predicted by Pierrehumbert to have allowed the phonotactic learning seen in Experiments 2A and 2B. In contrast, Bailey and Hahn's Generalized Neighborhood Model and Albright's Minimal Generalization Learner do not require any particular combination of token variability and/or type frequency for generalization to occur, so those models predict another generalization effect.

CHAPTER 6

EXPERIMENT 3: PHONOTACTIC GENERALIZATION FROM MULTIPLE WORD-TYPES

The first three experiments investigated what children could learn about a medial consonant cluster from a brief exposure accompanied by phonetic variation in the form of multiple talkers. In Experiment 1, phonetic variability was manipulated for a single familiarization word but did not influence how children produced that cluster in a new word during the test phase. In contrast, Experiments 2A and 2B combined token variability with type frequency, or multiple words, which did influence children's productions of the target clusters. In both experiments children's productions of Low English Frequency (LoEng) clusters were better when they first heard those clusters in three words spoken by multiple talkers.Based on these results, we have evidence that perceptual learning can lead to generalizations about phonotactics, provided that the multiple words support the generalization. Combined with the null result from Experiment 1, the results also provided strong support for the type-based phonotactic learning models proposed by Albright (2007) and Pierrehumbert (2003b). Both models propose that phonological knowledge, such as knowledge of phonotactic probabilities, is arrived at based on generalizations over unique word-forms. In other words, phonology is built from a word-type-based lexicon.

It is in their assumptions about the nature of the lexicon and the relevance of phonetic variation that the Albright and Pierrehumbert differ. Albright's model, the Minimal Generalization Learner (MGL), learns from a lexicon of abstract binary features of the kind proposed in Chomsky and Halle (1968), whereas Pierrehumbert's lexicon is a set of abstract word-forms learned with the aid of phonetic variability. Albright's model does not learn with reference to phonetic detail. Rather, word-forms are phonological feature bundles lacking in phonetic detail. In contrast, Pierrehumbert's model *requires* that word-forms be learned from variegated phonetic instantiations, such as word-tokens produced by different talkers. Furthermore, abstract word-forms must be learned *prior to* phonotactic learning, because phonotactics can only be learned from the abstract word-forms, and not directly from phonetic experience (cf. Figure 2.5 from Chapter 2, repeated below as Figure 6.1).

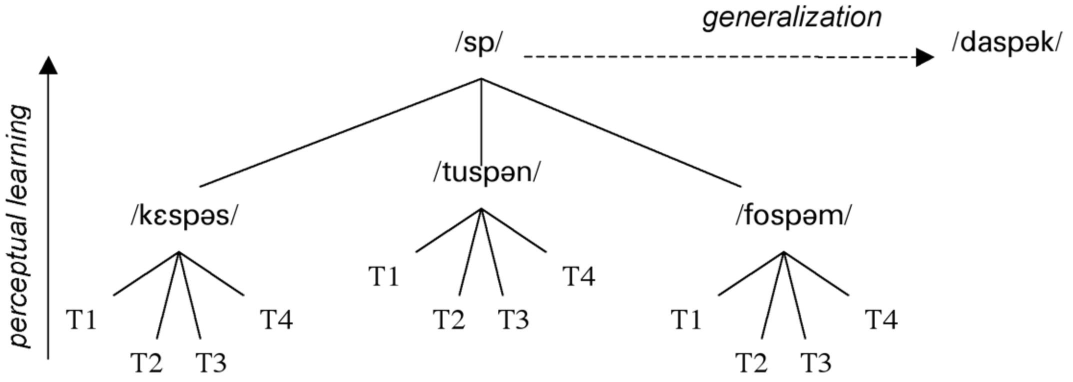


Figure 6.1: Creation of an abstract representation of the phonotactic sequence /sp/ according to Pierrehumbert's hybrid model.

With the properties of these two models in mind, we can see that they are telling different stories about why the results have come out as they have thus far. Albright's model treats the effects of perceptual learning in Experiments 2A and 2B as relevant to the type frequency manipulation only. The presence of token variability in those experiments was simply a distractor, possibly a detractor(cf. results showing word recognition times are slowed by changing talkers, reviewed in Goldinger, 1996) , from the process of learning phonotactics from the related words. Running the same experiment without token variability should result in similar, if not more robust, perceptual learning effects.

Pierrehumbert's model treats the results from Experiments 2A and 2B as dependent upon token variability, which is what allows abstraction of word-forms and, in turn, allows phonotactic learning. Without the token variability, the learning effects from those experiments should disappear. The different predictions made by the Albright and Pierrehumbert models are tested in the present experiment.

Children participating in Experiment 3 were familiarized with the same word-medial clusters used in the previous experiments, and they were asked to produce those clusters in a new word in a production test. As in Experiments 2A and 2B, half of the clusters were presented during the familiarization phase in three words (ExpHi), and the other half were presented in just one word (ExpLo). The same word sets used in Experiment 2B were used here. In the previous experiment, however, each familiarization word was spoken by four different talkers. In the present experiment, each of the four tokens of a word was spoken by the same talker. The total number of talkers, word-tokens, and word-types, however, was held consonant. The critical difference between Experiments 2B and 3, then, is that Experiment 2B contained intra-word talker variation, whereas Experiment 3 contained only inter-word talker variation.

The predictions made by all three models, Albright's model, Pierrehumbert's model, and the exemplar model are given in Table 6.1. Both the exemplar-based model and Albright's model predict that children should be able to generalize about a phonotactic sequence based on multiple word-types alone. The exemplar model, which tracks both type and token frequencies, predicts that children will create associations between the various tokens of related words, and, given the large number of tokens sharing ExpHi clusters, produce those clusters more accurately during the production test compared to the ExpLo clusters, for which only four tokens of one word were heard during the familiarization. Albright's model, which learns based on abstract phonological features, predicts that children will pull out the features common to the ExpHi words, that is, the features comprising the clusters, and use that knowledge to make more accurate productions of those clusters during the production test. Pierrehumbert's model uniquely predicts that children should not be able to generalize because, without phonetic learning, phonotactic learning should not be possible.

Table 6.1: Predictions made by the three models—the exemplar model, Pierrehumbert's hybrid model, and Albright's MGL—for the word-types-only experiment. *Gen* is given in cells where a model (row) predicts a significant effect of the Experimental Frequency manipulation for a given experiment (column). *No Gen* refers to a predicted null effect. A prediction that is unique to a particular model is boxed and in bold.

	EXP 1 Tokens	EXP 2A Types + Tokens	EXP 2B Types + Tokens	EXP 3 Types
EXEMPLAR MODEL Bailey and Hahn (2001)	**Gen**	Gen	Gen	Gen
HYBRID MODEL Pierrehumbert (2003b)	No Gen	Gen	Gen	**No Gen**
MINIMAL GENERALIZATION LEARNER Albright (2007)	No Gen	Gen	Gen	Gen

6.1 Method

6.1.1 Participants

Twenty-five children participated in this experiment. All were between the ages of 4;0 and 4;3 (mean = 4;0.27) and were recruited with the same database used for the previous three experiments. All children were native English speakers and had minimal exposure to other languages (average exposure to a second langauge was 0.7 hours per week, the mode was 0 hours per week). Only five children had regular exposure to another language, but their total weekly exposure was always less than 10 hours. Four children had some exposure to Spanish and one child had some exposure to ASL. Five children were reported by their mothers to have a personal or family history of speech, language, or cognitive deficits. These children were removed from the analysis. Four other children were removed from the analysis due to inattention to the directions or for not completing the experiment. The results reported below are for the remaining 16 participants, 8 girls and 8 boys.

6.1.2 Materials

The materials used in Experiment 3 are identical to those used in Experiment 2B. The familiarization words are repeated in Table 6.2. The target phonotactic sequences were again the medial clusters of CVCCVC nonsense words. The words were created and arranged such that the phonotactic probability of the cluster (English Frequency) and the number of words containing a cluster (Experimental Frequency) were manipulated. With respect to the English Frequency factor, half of the clusters had high phonotactic probabilities (HiEng) and half had low or zero probabilities (LoEng). With respect to the Experimental Frequency factor, half of the clusters were heard in three words during the familiarization (ExpHi) and half were heard in one word (ExpLo). The word sets were arranged to remove potentially distracting generalizations related to the word onsets. Additionally, the individual phone and biphone probabilities of the words, as well as the phone and biphone sums that comprised the words, were equated for the ExpHi and ExpLo word sets.

English Frequency

Individual phone and biphone scores, as well as phone and biphone sums, for the familiarization words are given in Appendix C. An ANOVA conducted on the two familiarization words sets and the test set found no significant differences between the three sets, but did find significant differences between the HiEng and LoEng words for phone and biphone scores related to the medial consonant cluster. Additional details about the analysis can be found in Section 5.1.2 of Chapter 5. The results confirmed that the word sets were reasonably well equated and that there were significant differences associated with the English Frequency factor.

Syllable Structure The English Frequency factor corresponds to differences in the permissibility of the clusters in syllables, as well as to differences in their phonotactic probabilities. The English Frequency × Experimental Frequency interaction found in Experiments 2A and 2B could be the result of differences in syllable structure, but it is also

Table 6.2: Familiarization words used in Experiment 3. The clusters are arranged according to their appearance in one of two experimental lists and one of two experimental blocks (a *block* refers to a familiarization phase followed by a test phase, with two blocks per experimental session). This is the same arrangement used in Experiments 2A and 2B. Items are written in IPA.

	ENG FREQ	**ITEM LIST 1**			**ITEM LIST 2**		**EXP FREQ**
Block One	**HiEng**	**mp**	**kt**		**sp**	**st**	*ExpHi*
		dɪmpət	lɛktəf		kɛspəs	mæstəm	
		nʌmpəs	saʊktəs		tuspən	baɪstəm	
		gumpən	biktəm		fospəm	gɪstək	
	LoEng	**ʃp**	**fp**		**mk**	**pk**	*ExpLo*
		foʃpəm	mæfpəm		nʌmkəs	saʊpkəs	
Block Two	**HiEng**	**sp**	**st**		**mp**	**kt**	*ExpLo*
		fospəm	mæstəm		nʌmpəs	saʊktəs	
	LoEng	**mk**	**pk**		**ʃp**	**fp**	*ExpHi*
		dimkət	lɛpkəf		kɛʃpəs	mæfpəm	
		nʌmkəs	saʊpkəs		tuʃpən	baɪfpəm	
		gumkən	bipkəm		foʃpəm	gɪfpək	

possible that the interaction reflects a ceiling effect, which may have obstructed the effect of Experimental Frequency on the HiEng clusters.

For continuity, the same clusters were used in Experiment 3. The expectation was that the presence or absence of another interaction and/or another ceiling effect would provide additional evidence about which of these explanations best captures the results of the previous experiments.

Experimental Frequency

Experimental Frequency was a manipulation of the number of times that a cluster was heard during the familiarization. ExpHi clusters were heard in three words, ExpLo clusters were only heard in a single word. Each cluster appeared as both ExpHi and ExpLo in the experiment by means of two word lists, which are listed in Table 6.2. Half of the participants heard one list and half heard the other. This is the same division used in Experiments 2A and 2B. Furthermore, the word lists also eliminated concerns about associations between words appearing in the same list, such as the presence of unwanted generalizations in the initial CVs or online neighborhood effects due to similar words. More details about the creation and arrangement of words can be found in Section 5.1.2 of Chapter 5.

Talkers Unlike Experiments 2A and 2B, the same talker produced each of the four tokens of each words heard during the familiarization phase. However, each word appearing in an experimental block was produced by a different talker. A different talker was assigned to the six ExpHi words (2 ExpHi clusters × 3 words) and and the two ExpLo words (2 ExpLo clusters × 1 word), so that all eight words appearing in an experimental block were produced by a different talker. This meant that the total number of talkers heard was the same in Experiments 2B and 3. It also ensured that the level of attention associated with hearing multiple talkers was the same across experiments.

The same eight talkers recorded for Experiment 2B (all native speakers of a western US dialect of American English) were used for the Experiment 3 words, and the same recording setup was also used. Talkers were asked to produce each word with trochaic stress and to articulate clear phonetic cues for both consonants in the medial cluster. A representative token of each word was chosen for use in the experiment, and all tokens were equated for RMS amplitude.

Familiarization Word Lists

The two lists shown in Table 6.2 were reversed so that words that had appeared in the second block appeared in the new lists in the first block. The full set of four lists (1A, 1B, 2A, and 2B) are given in Appendix A.

Test Words

The test words were identical to those used in Experiment 2B and are repeated in Table 6.3. The individual phone and biphone scores and the phone and biphone sums for entire words of the test words were matched with the scores and sums of the familiarization words sets so that all three word sets, ExpHi words, ExpLo words, and the test words, were all relatively similar with respect to phonotactic probabilities. See the description of the English Frequency factor in Chapter 5 for more details.

Table 6.3: Test words used in Experiment 3. These are the same test words used in Experiment 2B, and are matched with the familiarization words for individual phone and biphone scores, as well as for phone and biphone sums. Items are written in IPA.

High English	**Low English**
sp /daspək/	**ʃp** /daʃpək/
st /neɪstən/	**fp** /neɪfpən/
mp /sæmpəf/	**mk** /sæmkəf/
kt /tuktən/	**pk** /tupkən/

6.1.3 Procedure

The procedure for Experiment 3 was identical to that used for the previous experiments.

6.1.4 Analysis

The accuracy and production latency dependent measures from the previous experiments were again used in the analysis of Experiment 3. Accuracy scores were determined by two independent coders who worked together to eliminate inconsistencies and create a set of transcriptions for which there was 100% agreement. Disagreements accounted for less than 2% of the total data, and never more than three items were removed due to disagreement per participant. Production latencies were determined using Praat software (www.praat.org) by finding the time from the end of the target to the beginning of the child's own production. The coders were within 50 ms of each other for 90.0% of the

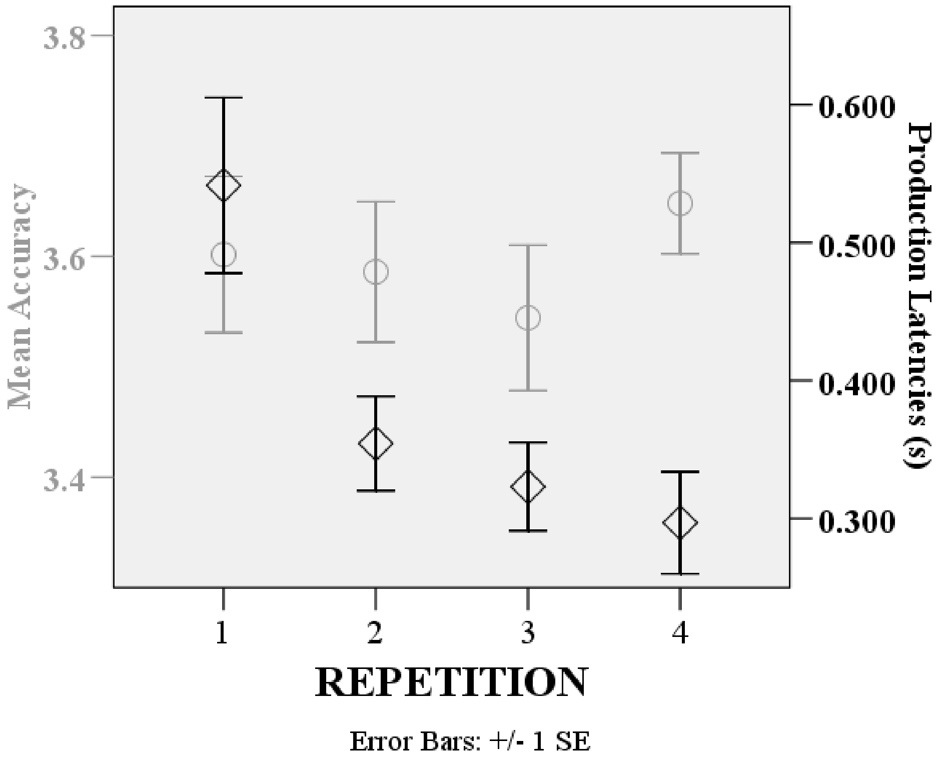


Figure 6.2: Average accuracy and production latency scores for repetitions 1-4 from Experiment 3. The accuracy averages are represented by open circles and are shaded gray. The black diamonds are the averages for the production latencies. The scale for the accuracy averages is on the left, the scale for the production latency averages is on the right.

latencies and within 100 ms for 95.3% of the latencies. Approximately 5% of the data was removed because children attempted the production two or more times.

As in the previous experiments, the first repetition of each word was removed from the production latency analysis. These repetitions were removed due to their high level of variance. All four repetitions were included in the accuracy analysis, however. Figure 6.2 gives the average accuracy and production latency scores for each repetition collapsed across subjects, items, and experimental factors.

6.1.5 Results

Table 6.4 gives the mean accuracy and production latency scores, as well as the standard deviations, for each of the four conditions in Experiment 3.

Table 6.4: Mean accuracy (out of 4) and production latency (in seconds) scores for Experiment 3. Standard deviations are given in parentheses.

Accuracy				
	HiEng *ExpHi*	*HiEng* *ExpLo*	*LoEng* *ExpHi*	*LoEng* *ExpLo*
Mean	3.76	3.793	3.307	3.421
(Std Dev)	(.284)	(.225)	(.561)	(.405)

Production Latencies				
	HiEng *ExpHi*	*HiEng* *ExpLo*	*LoEng* *ExpHi*	*LoEng* *ExpLo*
Mean	.321	.307	.317	.350
(Std Dev)	(.148)	(.119)	(.145)	(.153)

Averages of both dependent measures for each subject were entered into a 2 × 2 repeated measures ANOVA with English Frequency and Experimental Frequency as factors. For the accuracy analysis, there was a significant effect of English Frequency (F (1,15) = 42.269, $p < .001$, η_p^2 = .738). HiEng clusters were produced more accurately than LoEng clusters. The effect held for all 16 participants. There was no effect of Experimental Frequency (F (1,15) = .329, p = .575, η_p^2 = .021) and no English Frequency × Experimental Frequency interaction (F (1,15) = .177, p = .680, η_p^2 = .012). The results suggest that children were unable to learn anything from familiarization with multiple words sharing a medial consonant cluster when tokens of the familiarization words were presented without intra-word phonetic variation. A graph of the results is given in Figure 6.3.

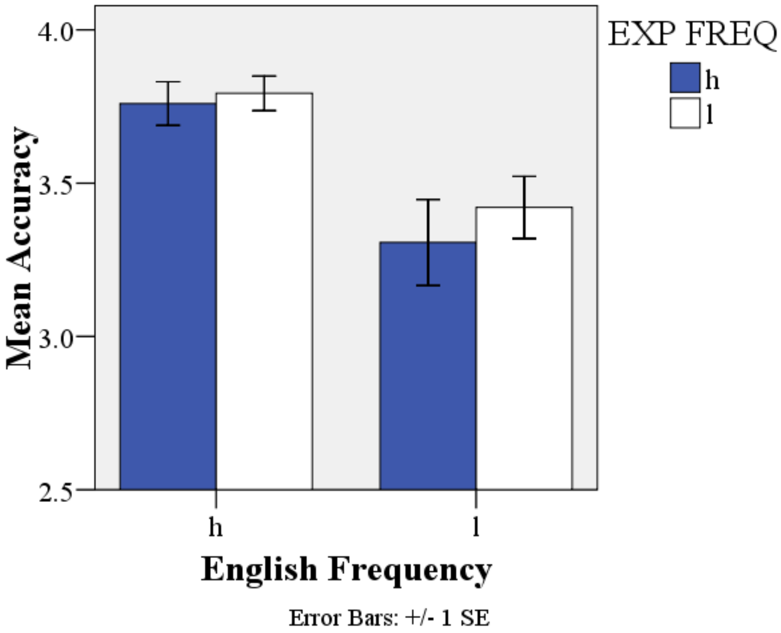


Figure 6.3: Bar graphs of the accuracy results from Experiment 3. On the ordinate is the mean accuracy for the medial consonant clusters, with a score of '4' being accurate production of both consonants. The bars are shaded according to Experimental Frequency and grouped by English Frequency.

For the production latency analysis, there were no significant main effects of English Frequency ($F = 1.028$, $p = .327$, $\eta_p^2 = .064$) or Experimental Frequency ($F = .184$, $p = .674$, $\eta_p^2 = .012$), and there was no English Frequency × Experimental Frequency interaction ($F = 2.647$, $p = .125$, $\eta_p^2 = .150$). It is not surprising that the production latency analysis resulted in a set of null effects, as this has been the trend for all four experiments. A graph of the results is given in Figure 6.4.

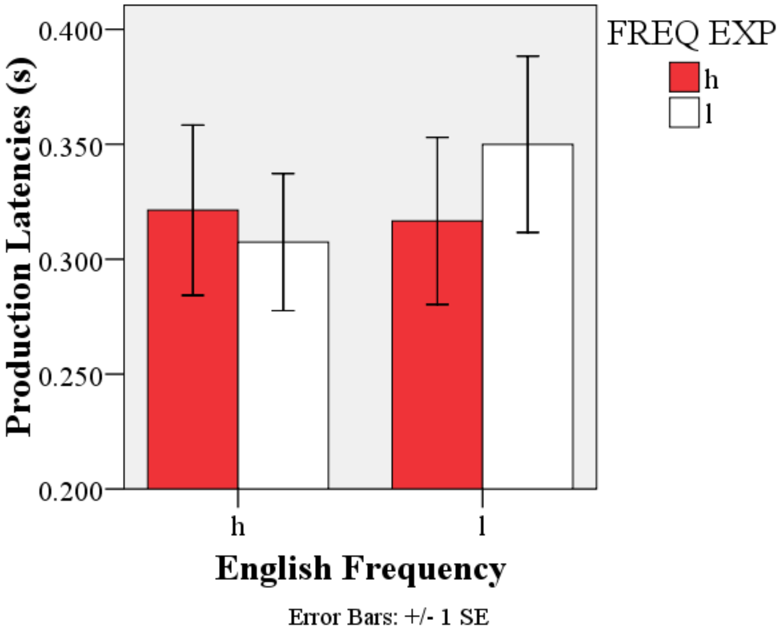


Figure 6.4: Bar graphs of the production latency results from Experiment 2B. On the ordinate is the time from the end of the target word to the onset of the child's production in seconds. The bars are shaded according to Experimental Frequency and clustered by English Frequency.

Ceiling Effects

Histograms of the accuracy scores are given in Figure 6.5. To determine whether there was a ceiling effect present in the accuracy measurement, average accuracy scores for each subject for the HiEng and LoEng clusters were analyzed using a Kolmogorov-Smirnov test for normality. There was a non-significant trend for the HiEng clusters (D (df = 16) = .197, p = .098) resulting from a reduced spread above the mean and a significant difference from a normal distribution for the LoEng clusters (D (df = 16) = .237, $p < .05$) resulting from reduced spread above the mean. It is possible given these results that effects of the Experimental Frequency manipulation were hidden for both LoEng and HiEng clusters, but visual inspection of the histograms suggests that the ceiling effect was far more prevalent for the HiEng clusters[1], and the significant difference from a normal distribution is likely caused by some other fact about the distribution of the data.

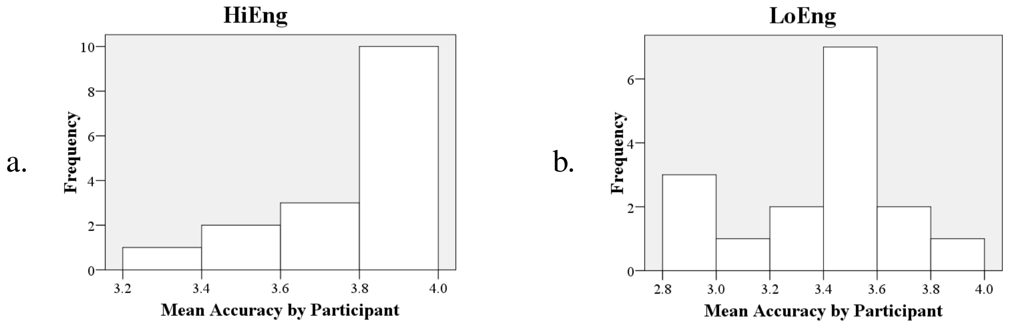


Figure 6.5: Histograms of the accuracy measurements for the HiEng and LoEng conditions in Experiment 3. Accuracy scores for the HiEng clusters are given in a. and scores for the LoEng clusters are given in b.

6.1.6 Discussion

As expected, children were more accurate when producing HiEng versus LoEng clusters. However, children's productions in Experiment 3 were not significantly affected by the

[1] The histograms clearly show that the majority of HiEng responses were at ceiling, but only a small number of LoEng responses were similarly high. This fact suggests that the test of normality does not adequately address the issue of *skewness* which is the most relevant statistic for addressing a ceiling effect. At the present time, I know of no statistical test that compares skewness across conditions. For this reason, it is important that future research addresses the ceiling effect in some way, such as recruiting younger participants. See Chapter 3, Footnote 10 for additional discussion.

manipulation of type frequency alone. In contrast to Experiments 2A and 2B, children were no more accurate to produce a cluster when they were familiarized with three related words compared to when they were familiarized with just one. The contrast can be explained by the difference in phonetic variability within words. In Experiments 2A and 2B, children heard each word spoken by multiple talkers. In Experiment 3, different talkers produced different words, but children heard each word produced by a single talker. Thus, the total amount of talker variability was held constant (at least for Experiments 2B and 3), but within-word variability varied from experiment to experiment, making it the primary source of explanation for the difference in results.

This contrast in results is similar to the difference between Experiments 2A and 2B and Experiment 1. In the latter experiment, token variation alone did not influence children's productions, which was accounted for by the lack of word-types to support the generalization. It appears, then, that neither type frequency nor token variability alone are effective in changing production accuracy. Instead, both variability at the phonetic level and type frequency at the word level are necessary for phonotactic learning.

Of the three models used to make predictions for Experiment 3, Pierrehumbert's (2003) hybrid model, which combines exemplar storage and abstract phonological structure, is uniquely capable of handling these results. This model proposes that the lexicon is a set of learned abstract word-forms built from a variable phonetic input, and that phonotactics are learned from those word-forms. Phonotactic learning is predicated on abstract word-forms and is not possible without them. Critically, abstract words are, "abstractions over phonetic space" (Pierrehumbert, 2003b, p. 179) and require some level of phonetic variability for the abstraction to take place. In Experiment 3, abstract word forms were not possible because abstract word forms could not be created from identical acoustic tokens. This is why Pierrehumbert's model predicted a null result, even when the target phonotactic pattern was present across multiple words.

Neither the exemplar-based model nor Albright's (2007) MGL is fully equipped to handle the results from Experiment 3. The exemplar model, which learns based on relationships between tokens, rather than relationships between unique lexical items, predicts that the consistency of the ExpHi clusters should allow learners to make associations between the various tokens, thereby influencing the productions of the test words that share that cluster. This prediction proved to be incorrect. Albright's model, on the other hand, learns from a lexicon composed of symbolic phonological features. The model does not address phonetic variability in relation to learning, although it could stipulate that variability is necessary for determining the featural makeup of a novel word. For the present, however, Albright's model appears to have incorrectly predicted that type frequency alone is sufficient for phonotactic learning.

One advantage of Abright's model, even in the face of the incorrect prediction, is that it provides a degree of specificity with respect to what phonological generalizations should look like beyond what can be gathered from either Pierrehumbert's model or the exemplar model. A case in point is the "island of reliability" effect discussed briefly in Chapter 2, that is, the finding that people are capable of tracking specific generalizations that might otherwise be subsumed under larger generalizations (Albright, 2002). Albright's model suggests that generalizations are often redundant, overlapping with other generalizations, and that language learners use that overlap to infer larger and larger generalizations. These higher-order generalizations provide a description of phonology that surpasses the scope of any existing exemplar model and the description offered by Pierrehumbert (2003b). Furthermore, it is beyond the scope of these results, which provide only a meager description of phonological structure. Rather, the present results are more relevant to the relationship between phonetics, or speech processing, and phonology. In particular, the results suggest that phonology is phonetics to some degree. That is, phonetic variability, which is generally considered to be outside the purview of phonology, has been shown to be a necessary component to biphone learning, and by extension, to phonological learning. How Albright's model, or any model of phonological learning, might incorporate this fact is an open avenue for future research.

Comparing the three models against the results from all four experiments, the hybrid model finds the greatest support. Unlike the exemplar model, the hybrid model predicted a null result in Experiment 1, in which token variability alone was manipulated to encourage phonotactic learning. No learning effect was found, validating the predictions of the hybrid and MGL models and creating significant difficulties for the exemplar model. All three models predicted learning in Experiments 2A and 2B, in which phonotactic patterns were familiarized using both token variability and type frequency, but only the hybrid and MGL models made the strong prediction that type-frequency was a *necessary* component to that generalization. Learning effects were found, providing further support for those models. Finally, Experiment 3 showed that type frequency alone did not promote any measurable learning effects, a result that only the hybrid model predicted. This is because the hybrid model requires phonetic variability to learn abstract word-forms and multiple word-types to learn phonological structure. When phonetic variation was missing from the familiarization words, type frequency did not affect children's productions, and it does not appear that children learned anything about the target clusters. This is exactly what we would expect to happen based on the hybrid model's structure.

The motivation for choosing Pierrehumbert's hybrid model was that the two other models did not correctly predict null effects. Incorrect predictions of null effects is a relatively weak form of evidence against a model, but the comparison of results across experiments allows for a degree of interpretation that might not otherwise be possible. The results of Experiments 1 and 3, considered in isolation, do little to help us understand which model is most capable of explaining how children learn phonotactics. When combined with Experiments 2A and 2B, however, Experiments 1 and 3 become useful because the provide a contrast with significant effects. We see that phonotactic learning is possible from a perceptual input, but only under certain conditions. The hybrid model was most sensitive to these conditions, and correctly predicted when children would show evidence of generalization and when they would not. Future research will improve upon these findings by the use of additional methodologies and sets of hypotheses in which all of the models make a unique set of both null and significant results. Nevertheless, the experiments presented here are rigorous in that they tested a set of predictions and selected an optimal model.

The results from the four experiments support the view that phonotactic learning proceeds from perceptual processes. Studies by Beckman and Edwards (1999); Edwards et al. (2004); Munson (2001) showed that child speech accuracy was correlated with phonotactic frequency, and Edwards et al. (2004) also showed that children's overall accuracy was correlated with the size of their lexicons, but these authors were only able to speculate about the mechanism responsible for these results. Results from the 10 Talkers study (Gerken et al., 2006), studies of the effects of phonetic variability on infant word recognition (Houston & Jusczyk, 2000, 2003; Singh, 2008), as well as studies of L2 learning (Bradlow et al., 1997, 1999; Flege, 1995; Wang et al., 2003), have all indicated that there is a linguistic payoff to perceptual learning from phonetic variability. In the case of infant word recognition, phonetic variability led to more robust word recognition that withstood unencountered forms of variability. In the case of the L2 learners, adult speakers showed improved L2 production because of exposure to variability. The 10 Talkers study showed that the same perceptual learning from variability also influences first language speech, as it improved the production speed and accuracy of four-year-old children.

Here we have seen that perceptual learning can also influence first language acquisition: what children learned about a consonant cluster through perception affected their productions of that cluster in a word that they had not encountered before. Beckman and Edwards (1999); Edwards et al. (2004); Munson (2001) showed that phonotactic probabilities were correlated with the production accuracy of new words, so children must have been using generalized knowledge of phonotactics in those studies. The studies I have presented show how that knowledge can come into existence. Children learn about phonotactics by tracking patterns in ambient speech. When those patterns are consistent across words, and the words have been encountered in a variety of phonetic environments, children learn phonotactics. Combined with the infant and L2 literatures, the scope of perceptual learning can be extended to phonological learning in general.

The combined results of Experiments 1-3 add to a body of evidence supporting the existence of abstract phonotactic knowledge. Claims about abstraction have been made by Albright in several places (Albright, 2002, 2007; Albright & Hayes, 2002, 2003), as well as by Pierrehumbert (2003a, 2003b, 2006), and is a conclusion reached by many others, as well (Bailey & Hahn, 2001; Beckman & Edwards, 1999; Bybee, 1995; Edwards et al., 2004; Goldinger, 2007; Hammond, 2003, 2004; Hay et al., 2003; Kisseberth, 1970; Pertz & Bever, 1975; Storkel & Morrisette, 2002; Vitevitch & Luce, 1998; Vitevitch et al., 1999). As discussed in the previous two chapters, abstraction is implicit in type-based learning, at least with respect to phonological learning, because types are defined with respect to abstract units such as phonemes or semantic categories[2]. Word-types were shown to be critical for phonotactic learning in Experiments 2A and 2B, so it is reasonable to conclude that abstract knowledge is also critical to phonotactic learning.

The importance of Experiment 3, particularly in comparison to Experiments 2A and 2B, is that it shows that abstract entities can be learned and that additional learning can take place on top of existing abstract forms. Children did not generalize in Experiment 3 because a phonetically variable input is necessary in order to learn word-forms. Children did not generalize in Experiment 1 because no generalization was available beyond the particular shape of each familiarization word. These facts are most parsimonious with an abstract representation of phonotactic probabilities, which are learned from abstract word-forms, which, in turn, are learned from a variable phonetic input. Remove either level and learning breaks down. This conditional learning is consistent with abstract learning that proceeds bottom-up, with higher-order learning dependent upon lower-level learning. If this is the correct view of the mental processes required for phonotactic learning, it is remarkable from the perspective of learning in general, whether phonotactic, phonological, or in any other domain.

In the next chapter, I combine the results from all four experiments in a single analysis to answer four questions that the individual experiments cannot answer themselves. First, I compare experiments with the same item sets to determine whether there are significant differences between those experiments. Second, I look at whether the results across all four experiments reflect a ceiling effect. Third, I compare the relative contributions of multiple repetitions and the Experimental Frequency manipulation in Experiments 2A and 2B, to assess how the results may be interpreted with respect to perceptual learning and articulatory practice. Fourth, I examine the behavior of individual clusters across all four experiments, which will provide some additional clarity about the role that syllable structure played in shaping the results. Following these additional analyses, I present some ideas for future research. The first proposal is a followup to Experiment 1 which

[2] I will return to the issue of how best to define type frequency in the next chapter. At this point, it is not clear whether types are best thought of as phonological entities, or semantic entities, or some combination of the two. Future research will be necessary to provide us with a better definition of this important concept.

equates the overall acoustic variability present in that experiment to the variability in Experiments 2A and 2B. The second proposed experiment considers what type frequency is from phonological and semantic perspectives. The third and final proposal looks at how perceptual learning might be used to better understand how phonotactics are related to sonority sequencing within the syllable.

CHAPTER 7

CROSS-EXPERIMENTAL ANALYSES AND FUTURE RESEARCH

In this chapter I address questions that may be answered by combining two or more of the experiments in a single analysis. These questions include 1) whether the cross-experimental comparisons made in previous chapters can be shown to be statistically significant, 2) the presence of a ceiling effect across the experiments, 3) the relative contributions of perceptual learning and articulatory practice to children's accuracy scores, and 4) some additional consideration of how syllable structure might have influenced the results. I also outline three future studies, including a modified version of Experiment 1, a study on how semantic content or reference play a role in type-based learning, and a study on the interaction of perceptual learning with the sonority hierarchy. I begin with a review of the results from the four Experiments discussed in Chapters 3 through 6, then set up the cross-experimental analyses, and finally move to the discussion of future research.

7.1 Review of Experiments 1-3

In Experiments 1, 2A, 2B, and 3, four-year-old children were familiarized with and tested on their ability to learn medial consonant clusters from a perceptual input. This was accomplished by requiring children to apply what they learned about those clusters in the familiarization to the production of new words containing the same clusters. Both familiarization and test words were CVCCVC with trochaic stress. The clusters were varied systematically for phonotactic probability: /sp/, /st/, /mp/, and /kt/ are common in English (i.e., they had high phonotactic probabilities), while /ʃp/, /fp/, /mk/, and /pk/ have low or zero-sum probabilities. This is the same manipulation of English phonotactic probability used in several related studies of child speech production (Beckman & Edwards, 1999; Edwards et al., 2004; Munson, 2001; Storkel & Rogers, 2000; Storkel, 2001, 2004; Zamuner et al., 2004).

The clusters also varied with respect to the number of times children heard the cluster, or Experimental Frequency. This was either a manipulation of token variation, of type frequency, or of both. Experiment 1 manipulated token variation alone, and children heard the familiarization words spoken by 1 or 10 talkers. Experiments 2A and 2B combined token variation and type frequency, and children heard all the familiarization words spoken by four talkers and heard one or three words for each cluster. In Experiment 3 type frequency alone was manipulated—children heard each familiarization word spoken by the same talker, but heard either one or three words per cluster. Importantly, the experiments tested children's ability to learn about the target clusters because the familiarization and test words were different, so children were required to use generalized knowledge of the clusters when producing the test words.

Measurements of production accuracy and production latency were made to determine how the English Frequency and Experimental Frequency factors affected children's speech. In Experiment 1, the token variability manipulation alone did not have a significant effect on children's productions. Children were no faster or more accurate to produce the target clusters in a new word when they first heard them in a word spoken by 10 talkers compared to when they heard them in a word spoken by a single talker. In Experiments 2A and 2B, however, the combination of token variation and type frequency did affect children's productions. Specifically, children were more accurate when producing the low English Frequency clusters if they first heard them in three familiarization words, and they heard each familiarization word spoken by four talkers. This pattern was consistent across both experiments, which used different item sets. No effects of either factor were found in the production latency measurement, however. In Experiment 3, there were again no differences in children's speed or accuracy, but in this case it was type frequency alone which appeared to have no effect. Consideration of all four experiments indicates that token variability and type frequency in isolation do not generalize, but the combination of the two does. This finding can be extended to explain why there are phonotactic probability effects in children's speech. As children hear and store phonetically variable word exemplars, they learn an abstract word-form that can then be combined with related word-forms to generalize about a phonotactic sequence. Thus, the difference in accuracy for HiEng and LoEng clusters reflects a variety of stored words and talker-tokens that support representations of the former group, but relatively few stored forms in support of the latter group.

7.2 Questions Arising from the Results of Experiments 1-3

This chapter addresses several questions that can be answered by comparing the results from all four experiments. Listed briefly, these questions are: First, are the differences between the different experiments themselves significant? Second, was there an overall ceiling effect, and if there was, which conditions were most affected? Third, how does the effect of perceptual learning compare to the effect of repeated repetitions? Does an analysis of repeated repetitions suggest phonotactic learning by articulatory practice? Fourth, should the English Frequency × Experimental Frequency interactions from Experiments 2A and 2B be attributed partially or fully to the differences in the syllable structure of the different clusters?

These additional analyses can be conducted by pooling the data from all four experiments into a single analysis. Pooling the data is possible because of the large degree of overlap across the experiments. The populations of each experiment were very similar, for example. All participants were four-year-old native speakers of American English with no personal or family history of speech, hearing, or language disorders. The same participants did not participate in multiple studies, which allows for a between-subjects analysis across the four experiments.

Second, similar materials were used for each experiment. The same clusters were used, and although one set of test words was used in Experiments 1 and 2A and another set in Experiments 2B and 3, all words had the same CVCCVC shape, were produced with clear phonetic cues for the clusters, and had trochaic stress.

With respect to the English Frequency and Experimental Frequency factors, which must be compared across the experiments, the similarity of procedure and analysis allow for these factors to be equated across experiments. The English Frequency manipulation was the same across the experiments because the same clusters were used throughout.

Experimental Frequency was always a manipulation of the number of times that a cluster was heard. All four experiments were composed of two blocks of familiarization followed by testing, and the dependent measures were always accuracy and production latency. The consistency of manipulations across the experiments, as well as the relative homogeneity of the participants and the similarity of the materials, allow for a cross-experimental analysis with Experiment as a between-subjects factor.

It should be noted that virtually all of the following analyses focus on the accuracy measurement. The production latency measurement was highly variable and consistently returned null results. This is in contrast to the significant effects found for the Experimental Frequency manipulation in the 10 Talkers experiment (Gerken et al., 2006). In that experiment, children were faster when producing a word if they first heard that word spoken by 10 different talkers during the familiarization. I was unable to reproduce this effect in any of the experiments discussed here, suggesting that the two experiments might probe different types of learning. There is some precedence for this claim. Munson et al. (2005) found that production latencies were suboptimal compared to other measures of phonotactic probability in child speech. It is possible that the task of word learning, as we might reasonably describe the 10 Talkers study, leads to a knowledge that affects the speech production system quite differently than the knowledge gained in a task of phonotactic learning. Regardless of whether this is the correct view of word learning and the production latency measurement, it is clear that latencies provided little information about the effects of the independent variables used here, so they will not be discussed further.

7.2.1 Cross-Experimental Comparison

The first question to be addressed is whether the differences between the experiments showing effects of Experimental Frequency are different than the experiments in which no effects were found. The designs of the individual experiments do not allow us to compare how token variability and type frequency combined affected accuracy compared to token variability and type frequency in isolation. By comparing the results across experiments, however, it is possible to determine whether children's overall performance when hearing token variability and type frequency was different than their overall performance when hearing only token variability or only type frequency.

Entering the four experiments in a single ANOVA is not the best method for answering this question, in part because one set of test words was used for Experiments 1 and 2A (first set), and another set was used for Experiments 2B and 3 (second set). The difference in test words turns out to affect overall accuracy significantly. Mean accuracy for the first set was 3.311 (*SD* = .311) and 3.611 for the second set (*SD* = .221). The average accuracy scores for the 16 participants in each experiment were compared in an ANOVA with Test Word Set as a factor and first set and second set as conditions. There was a significant difference in performance (F (1,63) = 19.811, $p < .001$, η_p^2 = .242) resulting from higher overall accuracy for the second set words. Therefore, a direct comparison of experiments with different word sets would not be particularly useful. The results of this analysis are presented in Figure 7.1.

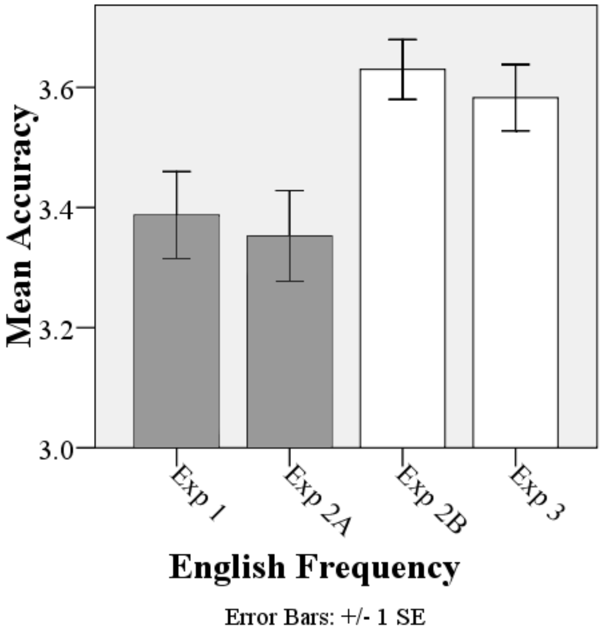


Figure 7.1: Average accuracy scores for the four experiments. Bars are colored according to the test word set that was used. The first set (Exps 1 and 2A) are colored gray, the second set (Exps 2B and 3) are colored white.

Although the four experiments could not be compared directly, experiments with the same word set were compared. Experiment 2A was be compared with Experiment 1 to determine whether overall accuracy was greater when children were exposed to clusters with both token variation and type frequency compared to token variation alone. The accuracy data from Experiments 1 and 2A were entered into a 2 × 2 × 2 ANOVA with Experiment as a between-subjects factor and English Frequency and Experimental Frequency as within-subjects factors. The three-way Experiment × English Frequency × Experimental Frequency interaction was significant (F (1,30) = 6.801, $p < .05$, η_p^2 = .185), which appears to be driven largely by the interaction from Experiment 2A. That is, Experiment 2A contained a two-way interaction but Experiment 1 did not. None of the two-way interactions were significant (Fs < 2), although the English Frequency × Experimental Frequency interaction approached significance (F = 3.033, p = .092, η_p^2 = .092) and appears to be driven by the English Frequency × Experimental Frequency interaction from Experiment 2A. With respect to main effects, there was a significant effect of English Frequency (F (1,30) = 52.612, $p < .001$, η_p^2 = .637), but not of Experimental Frequency (F (1,30) = .780, p = .384, η_p^2 = .025) and not of Experiment (F = .014, p = .749, η_p^2 = .003). A graph of the results is given in Figure 7.2.

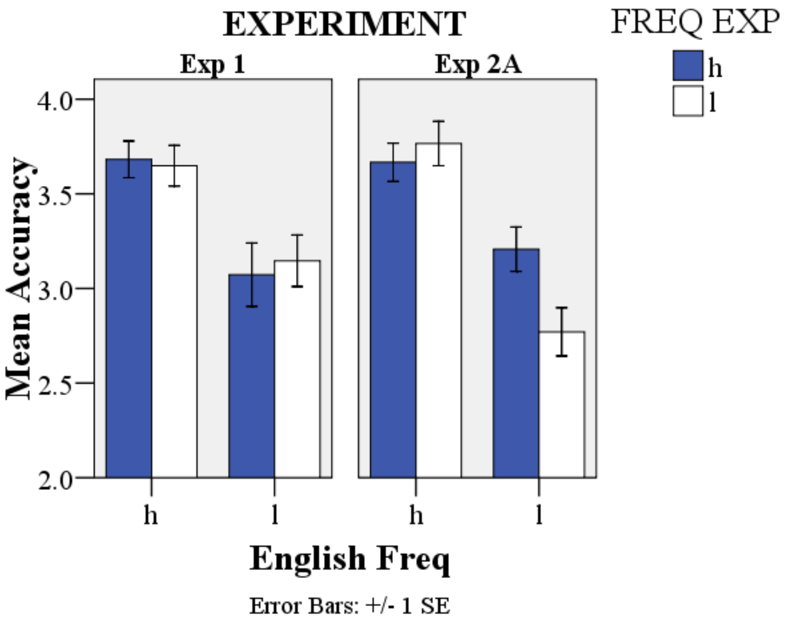


Figure 7.2: A comparison graph of the accuracy results for Experiments 1 and 2A. Experimental Frequency is coded by color, English Frequency by clustering, and Experiment by panel.

A related comparison was made for Experiments 2B and 3 to determine whether the combination of token variation and type frequency in Experiment 2B resulted in significantly different productions compared to the children in Experiment 3 that were only exposed to type frequency. The accuracy data from both experiments were entered into a single 2 × 2 × 2 ANOVA with Experiment as a between-subjects factor and English Frequency and Experimental Frequency as within-subjects factors. there was a significant three-way Experiment × English Frequency × Experimental Frequency interaction (F = 5.236, $p < .05$, η_p^2 = .124). As expected, this interaction resulted from the facilitative effect of the ExpHi condition for the LoEng clusters in Experiment 2B and a small decline in performance for the ExpHi LoEng clusters, as well as small declines in both HiEng and LoEng clusters in Experiment 3. There were no significant two-way interactions (all Fs < 2), but there was a significant main effect of English Frequency (F (1,30) = 78.920, $p < .001$, η_p^2 = .725). There was not a main effect of Experimental Frequency (F (1,30) = .113, p = .739, η_p^2 = .004) or of Experiment (F = .346, p = .561, η_p^2 = .011). A graph of the results is presented in Figure 7.3.

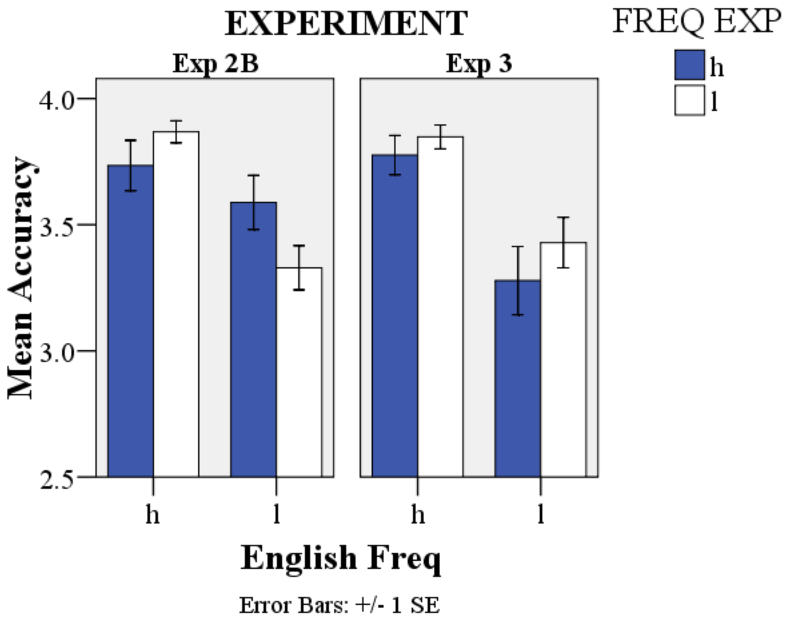


Figure 7.3: A comparison graph of the accuracy results for Experiments 2B and 3. Experimental Frequency is coded by color, English Frequency by clustering, and Experiment by panel.

The results of these analyses confirm that Experiments 2A and 2B were significantly different compared to Experiments 1 and 3, respectively. This, in turn, reinforces the claims made in previous chapters, particularly with respect to perceptual learning, that children only appeared to generalize the target phonotactic sequences when they were exposed to both phonetic variability and type frequency.

7.2.2 Ceiling Effects

The possibility of a ceiling effect was raised for each of the individual experiments. The purpose of this section is to explore where ceiling effects were occurring. Pooling the accuracy data from all four experiments, average performance was quite high. Several descriptive statistics, including the total number of data points, the mean, median, and skewness, are given in Table 7.1. The median value of ‘4’ means that more than half of the data consist of entirely accurate cluster productions.

The majority of completely correct cluster productions can also be seen in Figure 7.4,

Table 7.1: A table of descriptive statistics describing the accuracy results from all four experiments.

N	Mean	Median	Skewness
1877	3.44	4.0	-1.024

which shows histogram of the accuracy scores in the possible range of ‘1’ to ‘4’. Of the total number of data points collected, 86.1% were either a ‘3’ or a ‘4’. The histogram is contrasted with a normal distribution curve fit to the data. Given that the largest number of data points were scored as ‘4’, it appears that there is no part of the data corresponding to the right half of a normal distribution. This provides strong evidence that the accuracy scores were limited in their spread by a ceiling effect.

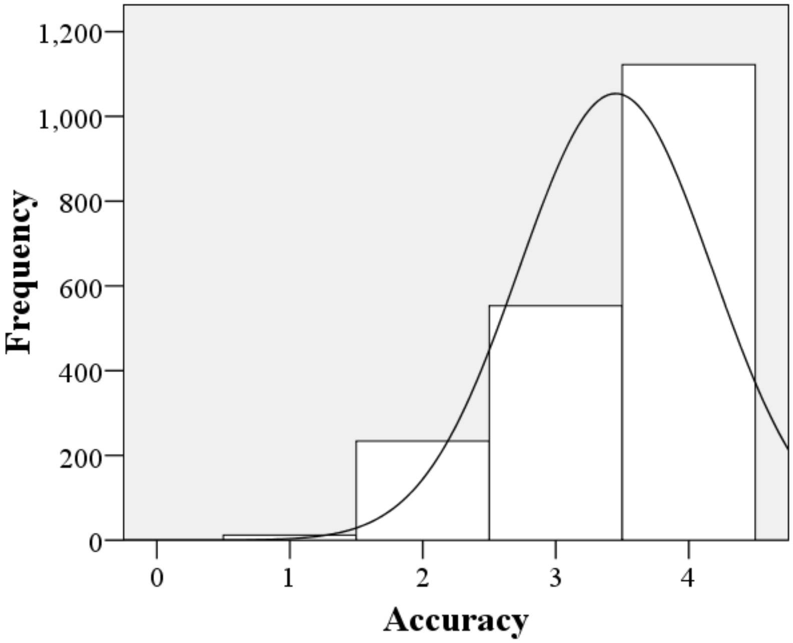


Figure 7.4: A histogram of the accuracy results from all four experiments.

In every experiment, the distribution of the data broken down by English Frequency suggested that the accuracy scores for the HiEng clusters were near or at ceiling. It is possible, then, that the effects of Experimental Frequency for the HiEng clusters were cut off by this ceiling. The range of scores from the individual experiments is shown in Figure 7.5. Boxplots of the data from each experiment show the spread from the minimum to the maximum score, and a division of the data into four quartiles. In every experiment, the maximum score was ‘4’, and the majority of scores are between ‘3’ and ‘4’. In fact, there is no upper quartile, or top 25% of the data, distinct from the upper-middle quartile for Experiment 2B. In Experiments 1, 2A, and 3, the range of the upper quartile is limited compared to the spread of the lower quartile.

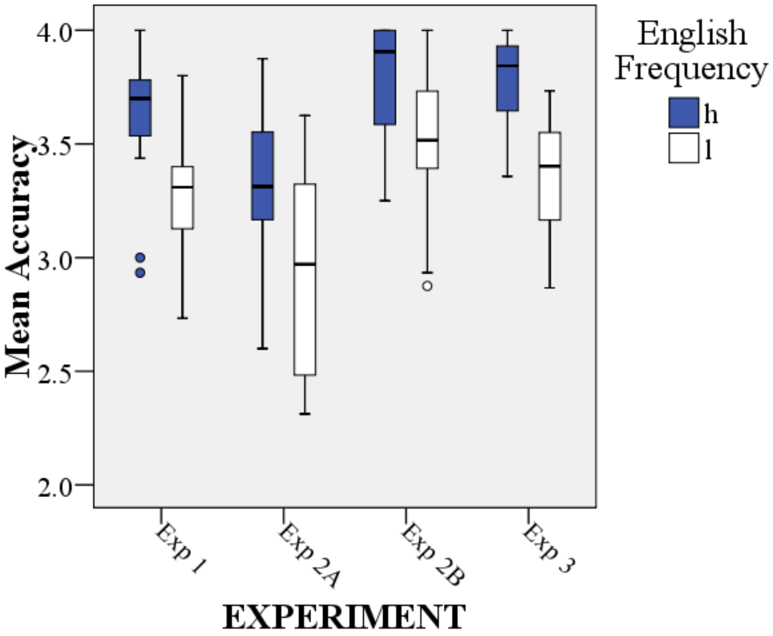


Figure 7.5: Box plots of the accuracy scores from the four experiments, split by the English Frequency factor. Data for the HiEng clusters apear in blue boxes; data for the LoEng clusters appear in white boxes.

Although there is clearly evidence of a ceiling effect, it appears to have influenced the HiEng cluster data more strongly than the LoEng data. This is important, because it affects the interpretation of both the individual experiments and the cross-experimental analyses. Consider, for example, the interaction in Experiments 2A and 2B, and whether the effect of Experimental Frequency can be called “facilitative.” Based on the interaction, tt does not appear that children’s productions of the HiEng clusters were facilitated by

Experimental Frequency. This may be explained by the ceiling effect, however. It is possible that children's representations of those clusters were sufficiently developed so that children were performing near perfectly, and the effects of the familiarization were hidden by the ceiling. This must be taken in the context of each experiment, as children in Experiment 2A scored much lower than children in Experiment 2B. As discussed in Chapter 4, however, the low scores can be attributed to difficulty with the item /simpən/, so it is still possible to interpret that data as reflecting a ceiling effect (cf. Figure 4.5). Future studies with younger children may help determine whether effects of Experimental Frequency can be found with these clusters.

The ceiling effect also influences the discussion of the three models. If a ceiling effect prevented the ability of token variability to generalize in Experiment 1 or type frequency to generalize in Experiment 3, then there is some chance that one or both of these models could have correctly predicted an effect of Experimental Frequency that was cut off by the ceiling. However, this claim is only relevant to the HiEng clusters, and does not change that fact that both models made incorrect predictions for the LoEng clusters. Given that the LoEng clusters do not appear to be limited by a ceiling effect, I do not believe that the ceiling effect changes how the experiments should be interpreted with respect to the predictions of the three models. It does, however, require future research to ensure that significant effects of Experimental Frequency were not hidden in Experiments 1 and 3, and to confirm that the criticism of the exemplar and MGL models is also relevant to HiEng clusters.

With respect to analyses below, the ceiling effect might have limited the degree to which perceptual learning and articulatory practice affected the data. Differential contributions of these types of learning were found and are discussed below, but it cannot be assumed that both factors were equally affected by the ceiling, so the conclusions must be considered tenuous. The ceiling effect is also relevant to interpreting the influence of syllable structure on the English Frequency × Experimental Frequency interactions found in Experiments 2A and 2B. It is tempting to think that syllable structure is in some way responsible for these interactions, but, as I will argue below, the ceiling effect provides a more obvious explanation, so role of syllable structure as it affects these data remains unclear. I turn now to each of these analyses.

7.2.3 Articulatory Practice vs. Perceptual Learning

A cross-experimental comparison can also address the relative effects of perceptual learning and articulatory practice. Recall that *articulatory practice* refers to the hypothesis that high production accuracy for high probability phonotactic sequences is the result of repeated articulations of those sequences, rather than the kind of perceptual learning motivated by the infant perception literature (Jusczyk, Friederici, et al., 1993; Jusczyk et al., 1994). Although the focus of the design here was on perceptual learning, it is possible to compare changes in accuracy from the first to the fourth repetition, as well. Figure 7.6 presents the repetition data averaged across subjects, words, and experiments. The general trend is for participants to be faster and more accurate with each subsequent repetition.

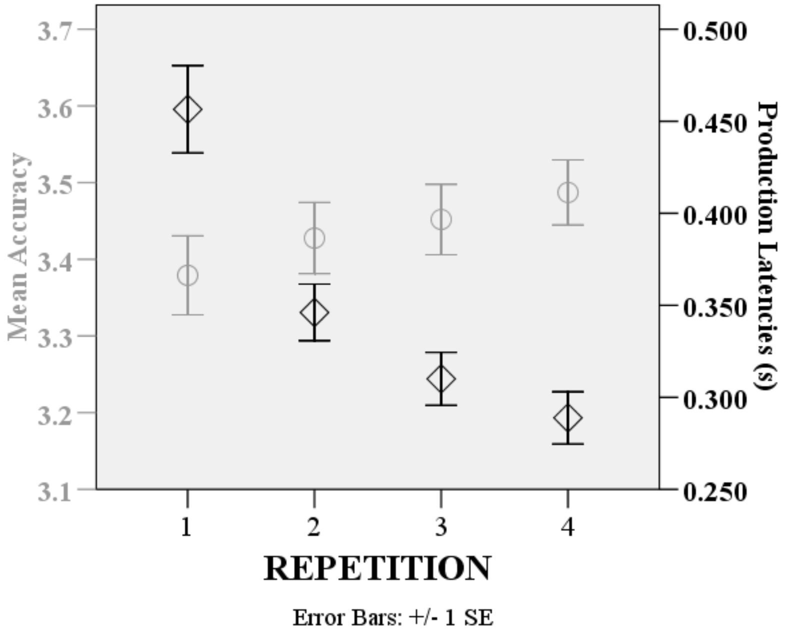


Figure 7.6: Changes in the accuracy and production latency measures across repetitions, collapsed across all four experiments. The accuracy averages are represented by open circles and are shaded gray. The black diamonds are the averages for the production latencies. The scale for the accuracy averages is on the left, the scale for the production latency averages is on the right.

To determine how strong these trends were, the data were entered into a within-subjects ANOVA with Repetition as a factor and repetitions ‘1’ to ‘4’ as separate conditions. There was a significant effect of Repetition for both the accuracy measurement (F (3,180) = 3.689, $p < .05$, η_p^2 = .058) and for the production latency measurement (F (3,180) = .633.398, $p < .001$, η_p^2 = .913). Of the two analyses of Repetition, it is clear that the larger effect was to speed up productions, as the effect size for the production latency analysis is much, much larger than the effect size for the accuracy analysis. However, this may not reflect improvements in articulation. Instead, it may reflect increased comfort with the task and an ability to predict the test words. Although the order of the test words for a given repetition was randomized, there were only four words, so children may have developed a level of expectation for the test word set and have been better prepared for producing the repetition ‘4’ compared to repetition ‘1’. The improvements in accuracy are more appropriately interpreted as reflecting improved productions, and they can be compared with the effect of Experimental Frequency (i.e., perceptual learning) from Experiments 2A and 2B.

Only Experiments 2A and 2B were chosen to represent the effect of perceptual learning because they are the only experiments where Experimental Frequency was found to have a significant effect, and therefore best reflect the overall influence of perceptual learning. Data from all four repetitions for Experiments 2A and 2B were chosen to represent the effect of articulatory practice to provide the largest and strongest data set for representing this factor.

To determine the relative contributions of Repetition and Experimental Frequency to the accuracy scores, the results were entered into a regression model. Often, a multiple linear regression model is appropriate for this type of analysis, but that is not the case here. The linear regression model assumes that the data are normally distributed, but the accuracy data are not. This is one of the consequences of the ceiling effect discussed above. Recall that Figure 7.4 shows how a histogram of the accuracy data compares with a normal distribution curve.

When a dependent variable is not normally distributed, the most appropriate regression model is a multinomial regression model, or a multinomial logit. The data from all four experiments were combined in the model. Repetition and Experimental Frequency were both entered as factors, as were English Frequency and two variables representing the two experiments where Experimental Frequency had an effect, Experiment 2A and Experiment 2B. Experiment 1 was used as comparison data for the variables representing Experiments 2A and 2B, but was not entered in the analysis. Experiment 3 was also omitted.

Each factor was entered into the regression equation independently so that the main effects of each variable could be compared. Unlike a linear regression, in which factors are compared by R coefficients, a multinomial regression uses the -2 log-likelihoods of a reduced model. The reduced model is produced by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0. The χ^2 statistic is the difference in -2 log-likelihoods between the final model and a reduced model.

In the final model there were significant contributions from English Frequency (χ^2 (3) = 184.287, $p < .001$), Experiment 2A (χ^2 (3) = 21.406, $p < .001$), Experiment 2B (χ^2 (3) = 33.217, $p < .001$), and Experimental Frequency (χ^2 (3) = 8.130, $p < .05$), but not of Repetition (χ^2 (9) = 13.760, $p = .131$). The contribution of English Frequency can be attributed to the fact that HiEng clusters were produced more accurately than LoEng clusters. The significant contributions of the two experiments can be attributed to the differences in accuracy for the different word sets, that is, children were generally more accurate in Experiment 2B than in 2A. The contribution of Experimental Frequency can be attributed to the learning effects seen in both experiments, particularly to the LoEng clusters. No significant contribution was made by the Repetition factor, however, suggesting that children's accuracy did not vary systematically as a result of repeated productions.

The results may be taken as validation for the initial choice of perceptual learning as a means for studying phonotactic learning. The significant contribution of Experimental Frequency to the regression model shows that perceptual learning had a significant effect on children's productions. This effect can be described as facilitative for low probability clusters, but may also have been hampered by the ceiling effect for the high probability clusters. The non-significant contribution of Repetition may also have been hampered by a ceiling effect, but the fact that it was not significant in the face of a significant

Experimental Frequency effect suggests that, in this study, its relevance to phonotactic learning appears to be less influential compared to perceptual learning.

The results of the multiple regression analysis may be extended to better understand the relationship between the effect of Experimental Frequency in this experiment and the broader reality of phonotactic probabilities. Probabilistic phonotactic effects such as those related to the English Frequency manipulation have been attributed to lexical statistics by a number of different sources (Albright, 2007; Bailey & Hahn, 2001; Coleman & Pierrehumbert, 1997; Greenberg & Jenkins, 1964; J. Ohala & Ohala, 1986; Pierrehumbert, 2003b). Phonotactic effects in production must arrive from perceptual learning, or articulatory practice, or some combination of the two. The fact that Experimental Frequency contributed to the regression model but Repetition did not suggests that, of these two sources for English Frequency effects, Experimental Frequency is the more relevant.

One question that the analysis above cannot answer is whether there are limits to either perceptual learning or articulatory practice. Future research might explore the maximum improvement that can be provided by each factor, and whether a combination of the two can be used to optimize speech accuracy. This would be a natural extension of the L2 studies conducted by Bradlow et al. (1997, 1999) and Wang et al. (2003), which focused exclusively on perceptual learning. A a study of how perceptual learning and articulatory practice combine to affect speech production would also have many clinical implications. Children in speech therapy are often asked to repeat difficult sounds, but little emphasis has been placed on how perceptual learning, particularly the combination of phonetic variability and type frequency studied here, might also improve disordered speech. See the "Clinical Implications" section of the review of phonotactic effects made by Storkel and Morrisette (2002) for a related proposal.

7.2.4 Syllable Structure

A third question arises from the differences in the syllable structure of the HiEng clusters compared to the LoEng clusters. All four HiEng clusters—/sp/, /st/, /mp/, and /kt/—are allowable as syllable onsets, or codas, or both; the four LoEnglish clusters—/ʃp/, /fp/, /mk/, and /pk/—are generally not allowed as either onsets or codas. Given the interactions in Experiments 2A and 2B, it is possible that perceptual learning is mitigated by the syllabic properties of the target clusters.

Section 3.1.2 of Chapter 3 provides a detailed discussion of how phonotactic probability and syllable structure are both potential contributers to the English Frequency variable. In that chapter, I argued that, in spite of the fact that both phonotactic probability and syllable structure could explain effects of English Frequency, several past studies (Munson, 2001; Edwards et al., 2004; Hay et al., 2003) have found effects of phonotactic frequency apart from syllable structure using linear regression models, so it is reasonable to interpret the results of these experiments with respect to phonotactics.

In those past studies, it was relatively easy to attribute accuracy scores to phonotactic probability because there was a linear relationship between the two: the higher the phonotactic probability, the higher the accuracy scores. In the case of Experiments 2A and 2B, the interaction of English Frequency and Experimental frequency makes this conclusion much less obvious. Why does hearing a LoEng cluster in three familiarization words lead to improved accuracy, but not so for the HiEng clusters? One possible answer is that the Experimental Frequency manipulation, and perceptual learning in general, is only relevant to certain types of syllabic units. For the LoEng clusters, which do not correspond to either licit onsets or codas, perceptual learning may provide a clear benefit. For the HiEng clusters, which are known and syllabifiable entities, perceptual learning may have no effect, as Experiments 2A and 2B suggest. Another way to phrase this is to say that perceptual learning is irrelevant to clusters which conform to English syllable structure, and so perceptual learning is not particularly relevant to English.

The most important answer to this concern is that a ceiling effect, rather than a difference in syllable structure, is the most likely explanation for the lack of an Experimental Frequency effect for HiEng clusters. As was demonstrated in Section 7.2.2 above, the HiEng clusters were consistently at ceiling. It is possible that Experimental Frequency would have behaved consistently with respect to both the HiEng and the LoEng clusters if there the scores had been allowed to spread upwards. Thus, it would be best to eliminate the ceiling effect before interpreting the interaction as evidence that perceptual learning is irrelevant to syllabifiable entitites. No solution to this problem can be given here, but future research with a younger age group, perhaps 3- or 3 1/2-year-olds, should eliminate the ceiling and provide us with a better understanding of where and when Experimental Frequency affects children's productions.

As a final note, recall that in Chapter 3 I argued that the confounding of phonotactic probability and syllable structure was an unavoidable outcome resulting from the division of phonotactic frequencies into discrete categories. The high probabilities of clusters like /sp/, /st/, /mp/, and /kt/ is entwined with their phonological status. It may be that their separation, both in experimentation and in theory, is an artificial distinction and does not reflect the underlying nature of these clusters. No good answer to this problem can be given here. Rather, the results suggest that a great deal more can be done to understand whether and how children learn both phonotactic probabilities and syllabic structure. It may be the case that the English Frequency × Experimental Frequency interaction reflects a fundamental difference in how different phonological units respond to a perceptual input, or it may be that the interaction was an artifact created by a ceiling effect. I take these different explanations to be evidence of the vitality of this line of research, rather than fundamental problems. In the next section, I explore to avenues for future research, one of which outlines an approach to how syllabic structure can be better understood using a perceptual learning paradigm.

7.3 Future Research

In this section, I outline a follow-up to Experiment 1, as well as two related and promising avenues for future research. One study will seek to add precision to the definition of "word type", a term that is ambiguous with respect to its phonological and semantic status. The second involves an examination of how perceptual learning interacts with an important constraint on the production of consonant clusters: the Sonority Hierarchy (Jespersen, 1904).

7.3.1 Equating Overall Phonetic Variability

The first experiment proposed for future research relates to the overall phonetic or acoustic variability present across the experiments. As noted in Chapter 4, Experiment 2A (as well as Experiment 2B) contains 12 ExpHi items compared to the 10 ExpHi items in Experiment 1. Although the ratio of variability in ExpHi and ExpLo conditions was greater in Experiment 1, it is possible that the absolute level of variability is responsible for the perceptual learning effects in Experiments 2A and 2B, rather than the added manipulation of type frequency. To address this concern, a follow-up to Experiment 1 can be conducted in which 12 talkers are used for the ExpHi condition. This will equate the level of overall acoustic variability present across Experiments 1, 2A, and 2B (although the acoustic variability in Experiment 3 remains significantly less than in any of the other three experiments), allowing for the elimination of absolute acoustic variability as an explanation for the differences between experiments.

7.3.2 What is a "Type"?

In the discussion of of type-based learning in Chapter 2, a great deal hinged on how type frequency was determined. This was especially true in the case of Bybee (1995), who disputed an earlier definition of type frequency given by Clahsen and Rothweiler (1992) that included inflected forms as unique types. But certain aspects of our understanding of type frequency appear to be missing in Bybee's description as well as in others (Albright & Hayes, 2003; Albright, 2007; Pierrehumbert, 2003b), and they persist in the studies presented here. The most pressing concern is whether "types" should be defined in semantic or phonological terms. This concern is felt acutely with respect to our understanding of abstraction, because it determines what we consider an abstract category to be.

Consider Experiments 2A and 2B, in which children were more accurate to produce clusters like /ʃp/ and /mk/ when they heard them in multiple familiarization words. Those words could be given type status for at least two reasons. One, they were phonologically distinctive words (e.g., kɛʃpəs, tuʃpən, and foʃpəm), that is, they were composed of different sets of phonemes. Two, they were assigned to different make-believe animals, giving them distinct semantic qualities. It is not clear which of these factors is most relevant to the creation of abstract categories. Additionally, we might ask whether both phonological and semantic status are necessary for the creation of an abstract-word form, just as phonetic variability and type frequency both appear necessary for phonotactic learning.

One approach to this question of “type” status is to systematically vary semantic and phonological status in a perceptual learning study such as those presented here. A study along these lines might vary how words are assigned to referents. A graphical illustration of this design is given in Figure 7.7. One condition would be virtually identical to the ExpHi conditions in Experiments 2A and 2B. Children hear multiple words that differ phonologically, are spoken by multiple talkers, and each have unique referents (cf. 7.7a). Based on the results presented here, we would expect to see generalization. In the other condition, children would hear those same words, spoken by multiple talkers, but given the same referent (cf. 7.7b). If the semantic component to a type is critical, children should generalize in the former condition but not in the latter. If the semantic component is not critical, we would expect to find generalization in both conditions.

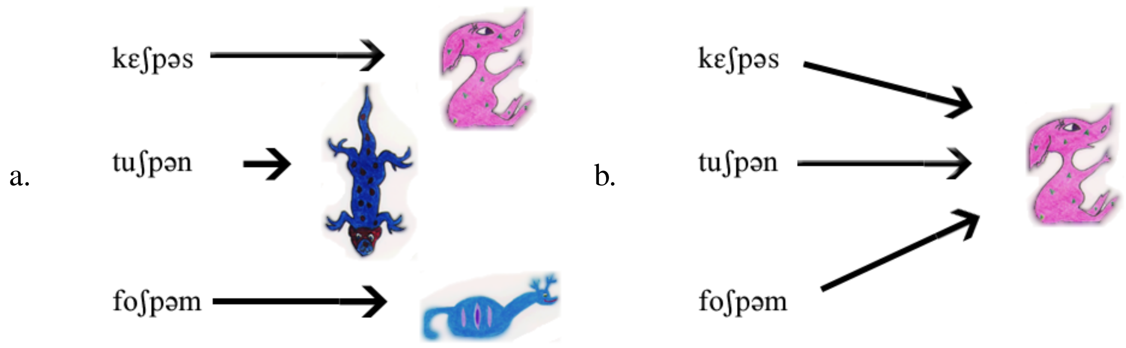


Figure 7.7: A comparison of “semantics” and “no semantics” conditions in a study of word types. In 7.7a, each word receives a unique semantic interpretation. In 7.7b, all words are given the same semantic interpretation.

7.3.3 Perceptual Learning and the Sonority Hierarchy

A second avenue for future research would be to explore how the sonority profile of the children’s errors relates to the Experimental Frequency factor. Children often drop a consonant when producing a consonant cluster, or they produce it in error. Work by D. K. Ohala (1996, 1999), for example, has shown that cluster reduction by very young children often conforms to predictions made by the sonority hierarchy (Jespersen, 1904). Nevertheless, English learners do eventually learn to accurately produce consonant clusters, an achievement still open to explanation (but cf. Bernhardt & Stemberger, 1998, for one proposal and related references). One possibility is that they learn to produce those clusters through the same type of perceptual learning seen here. This could be addressed using the present data by recoding children’s errors based on the sonority hierarchy[1]. Sonority-based errors could then be compared across ExpHi and ExpLo conditions. If perceptual learning plays an important role in learning to produce clusters accurately then we would predict fewer sonority-based errors for productions of ExpHi clusters.

[1] The precise details of the recoding could work in at least two ways. One possibility is to use data in which children completely dropped a consonant. Another possibility would be to code errors with respect to their similarity to onset or coda sonority contours. Given the relatively few number of consonant deletions in the data, this latter approach is more likely to prove fruitful

7.4 Conclusion

The additional analyses conducted across all four experiments suggest that many of the conclusions reached for individual experiments are relevant to the set of experiments as a whole. Most importantly, the individual ceiling effects were shown to reflect a general trend of ceiling effects for the HiEng clusters. This fact affected the interpretation of the other analyses, but does not refute the conclusions reached earlier. The interpretation of the English Frequency factor as reflecting differences in syllable structure, for example, appears to be more readily interpreted as the result of a ceiling effect. The ceiling effect may also have limited the degree to which both perceptual learning and articulatory practice contributed to the accuracy data, although the Experimental Frequency factor was still found to contribute significantly to the accuracy data, supporting the interpretation of Experiments 2A and 2B as indicative of perceptual learning. Finally, two related lines of research were proposed for the future, which may yield additional evidence for the conclusions reached here. They include a study on the semantic and phonological contributions to “type” status and a study of the relationship between perceptual learning and sonority. In the next and final chapter, I review the motivation behind the four experiments, the design and results, and I interpret these results within the broader scientific pursuits of phonology and language acquisition.

CHAPTER 8

GENERAL DISCUSSION AND CONCLUSION

8.1 Reviewing the Context, Purpose, and Relevance of the Dissertation

This dissertation has focused on the property of sound sequencing known as phonotactics, or the probability or likelihood that a sound sequence appears in a particular location in a word. Phonotactic probabilities are more than just facts about a language's words, however. They are facts about minds. People—adults, children, even infants—have a cognitive capacity for storing and utilizing phonotactic information. We know this because infants use phonotactics as cues to word boundaries (Mattys & Jusczyk, 2000; Mattys et al., 1999), children depend on phonotactics to learn new words (Storkel, 2001, 2004; Storkel & Morrisette, 2002; Storkel & Rogers, 2000), and adults use phonotactics when they make judgments about the naturalness of words they are encountering for the first time (Albright, 2007; Bailey & Hahn, 2001; Coleman & Pierrehumbert, 1997; Greenberg & Jenkins, 1964; Hammond, 2003, 2004; J. Ohala & Ohala, 1986). Finally, phonotactics are important because they tell us about the lexicon and about how words interrelate (Albright, 2007; Bailey & Hahn, 2001; Coleman & Pierrehumbert, 1997; Luce & Pisoni, 1998; Pierrehumbert, 2003a, 2003b; Vitevitch & Luce, 1998; Vitevitch et al., 1999; Vitevitch & Luce, 2004).

The purpose of this dissertation was to address several unanswered questions about what phonotactics are and how they are learned. First, I asked whether a perceptual learning mechanism could be responsible for the phonotactic effects seen in both infant perception and child speech. Perceptual learning is doubtlessly behind what infants know about phonotactics. With respect to child speech, previous studies have consistently found that children are more accurate when producing high probability phonotactic sequences (e.g., Beckman & Edwards, 1999; Edwards et al., 2004; Munson, 2001; Zamuner et al., 2004), but the results leave the learning mechanism behind the effect unexplained. To address this gap in our understanding, I used an artificial language learning paradigm (Gerken et al., 2006) to create phonotactic effects within the experimental setting. By showing that perceptual learning could lead to differences in speech production accuracy, the experiments provided a means for testing whether perceptual learning is behind the phonotactic probability effects characteristic in normal speech development.

Second, I asked whether the phonotactic representations that children formed were dependent on learning from phonetic variability, or type frequency, or both. Phonetic variability has been found facilitate word learning in infants (Houston & Jusczyk, 2003; Singh, 2008), the production of non-native contrasts in second language acquisition (Bradlow et al., 1997, 1999; Wang et al., 2003), and novel word production in young children (Gerken et al., 2006). Type frequency has been implicated in studies of phonological patterning, including morphological correspondences (Albright & Hayes, 2003; Bybee, 1995; Pierrehumbert, 2003b) and phonotactic probabilities (Albright, 2007; Pierrehumbert, 2003b). By asking whether phonetic variability and type frequency are important in child speech development, several theoretical consequences ensued. Most importantly, the respective contributions of types and tokens to phonotactic learning were shown to be important components of three models of phonological acquisition.

The three models were an abstraction-free exemplar model, the Generalized Neighborhood Model, described by Bailey and Hahn (2001), a combined exemplar + abstraction, or hybrid, model proposed by Pierrehumbert (2003b), and an abstraction-only model, the Minimal Generalization Learner (MGL), detailed in Albright (2007). Each makes a unique set of predictions with respect to the contributions of word-tokens and word-types to phonotactic learning. Bailey and Hahn's (2001) Generalized Neighborhood Model does not really learn phonotactics. It stores word-tokens and associates them based on shared phonemes. The simultaneous activation of a set of words that share a particular property, such as a medial cluster, is how the model arrives at phonotactic effects. Pierrehumbert's (2003b) model is also based on the storage of word-tokens, but learns abstract representations from statistically robust patterns across those tokens. From low level abstractions, Pierrehumbert's model learns progressively more abstract representations. For example, the model learns abstract word-forms from phonetic experience, then learns phonotactics from the abstract word-forms. Albright's (2007) MGL stores words as symbolic feature bundles, so it doesn't learn from phonetic experience or word-tokens. Instead, it learns generalizations about the feature sets of two or more words and stores those generalizations separately.

The predictions of these three models are discussed in greater detail in previous chapters, particularly Chapter 2, so I will only describe them here briefly. The exemplar model is the only model that predicts that a phonotactic pattern present across multiple word-tokens should be able to generalize to a new word. This is because, to the extent that a phonotactic pattern is consistent across phonetically separable word-tokens, the model should make the associations between these word-tokens necessary to "learn" the relevant phonotactic pattern. In contrast, the Albright and Pierrehumbert models predict that phonotactics must be learned from word-types, and that word-tokens alone will not generalize. Finally, the Pierrehumbert model is the only model to predict that a combination of word-tokens and word-types are necessary for generalization to occur. This is because Pierrehumbert's model relies on abstract word forms learned from a variable phonetic input, whereas Albright's model is concerned only with the abstract properties common to a set of words.

8.2 A Summary of the Experiments and the Results

To test the models above, I used a methodology described by Gerken et al. (2006) in their 10 Talkers study. The design owes a great deal to the infant artificial language learning literature (cf. Gómez & Gerken, 2000 for a review). As in the 10 Talkers study, the experiments presented here involved an initial artificial-language or familiarization phase followed by a speech production test. Children were familiarized with a series of nonsense words that contained a target medial consonant cluster. To determine whether those manipulations led to some form of learning, children were tested on their knowledge of the medial cluster in a production task. They were played a new word containing the target cluster which they were asked to repeat. Unlike infant studies, which measure looking times related to the test words, these experiments measured children's production speed and accuracy. The advantage of this design was that it allowed learning to be measured in the same way that previous studies have measured the effects of phonotactic probabilities (cf. Beckman & Edwards, 1999 and subsequent related studies). Furthermore, it allowed for a determination of which particular qualities of the familiarization words, such as token variability and type frequency, were most important for supporting learning.

The target phonotactic sequences in these experiments were the medial consonant clusters of CVCCVC nonsense words with trochaic stress. Following previous studies of phonotactic probability effects in child speech, the English Frequency of those clusters was varied systematically. The clusters /sp/, /st/, /mp/, and /kt/ were common in English, while the clusters /ʃp/, /fp/, /mk/, and /pk/, had low or zero-sum probabilities. The expectation was that children would produce the high probability clusters most accurately. For assessing perceptual learning, however, the critical manipulation was of the frequency and quality of the exposure to those clusters within the experiment, or Experimental Frequency.

Experimental Frequency was a manipulation of how often children heard the target clusters and was meant to simulate the differences in ambient language exposure that may underly English Frequency effects. In the ambient language of English learners, high probability phonotactics are heard often, whereas low probability phonotactics are heard infrequently or not at all. The experimental manipulations of token variability and type frequency were meant to reflect the fact that children hear high probability clusters more often and in more words compared to low probability clusters.

With respect to token variability, the familiarization tokens were spoken by either one or many talkers. Gerken et al. (2006) had shown that children are more accurate when producing a word like /mæfpəm/ if they first heard it produced by 10 talkers during the familiarization. I wanted to know whether token variability, instantiated as multiple talkers, would allow children to generalize the medial consonant cluster of a word like /mæfpməm/ to a new word like /neɪfpən/.

With respect to type frequency, the cluster tokens were associated with either one or three words. Numerous authors (Albright, 2007; Albright & Hayes, 2003; Bybee, 1995, 2001; Pierrehumbert, 2003a, 2003b) have suggested that the existence of distinct lexical items that share phonological properties, that is, type frequency, is the key ingredient to learning phonotactic patterns. To test this, I looked at whether hearing a cluster in three familiarization items, such as /mæfpəm/, /baɪfpəm/, and /gɪfpək/, would have a greater influence on children's productions of /fp/ than would hearing one word, such as /mæfpəm/.

Over the course of four experiments, I manipulated token variability and type frequency independently to determine whether one, the other, or the combination of the two, was most influential in allowing children to generalize the phonotactic patterns. Experiment 1 manipulated token variability alone—children heard the familiarization words spoken by 1 or 10 talkers. Experiments 2A and 2B combined token variability and type frequency—children heard all the familiarization words spoken by four talkers and heard one or three words for each cluster. Experiment 3 manipulated type frequency alone—Children heard each familiarization word spoken by the same talker, but heard either one or three words per cluster. With respect to all four experiments, the three models of phonotactic learning made different predictions. The exemplar model, most sensitive to tokens, predicted generalization in all of the experiments because all of the experiments contrasted high and low token variability. Albright's Minimal Generalization Learner, most sensitive to type frequency, predicted generalization in Experiments 2A, 2B, and 3, in which type frequency was varied. Pierrehumbert's hybrid model, which is sensitive to both token variability and type frequency, only predicted generalization in Experiments 2A and 2B.

Two dependent measures of the children's productions were collected: an accuracy score based on transcriptions, and the time from the end of the target word to the beginning of the child's production, or the production latency. Only the accuracy measurement was found to be sensitive to the English Frequency and Experimental Frequency factors, however, so discussion of the results focus on the accuracy analysis. In Experiment 1, the token variability manipulation alone did not have a significant effect on children's productions. Although there was an effect of English Frequency—children were more accurate when producing the high probability clusters—children were no more accurate to produce the target clusters when they first heard them in a single familiarization word spoken by 10 talkers compared to when they heard them in a familiarization word spoken by a single talker. In Experiments 2A and 2B, however, the expected effect of English Frequency was accompanied by an interaction of English Frequency and Experimental Frequency. Children were more accurate when producing the low English Frequency clusters if they first heard them in three familiarization words and they heard each familiarization word spoken by four talkers. This pattern was consistent across both experiments, and provides evidence that perceptual learning can lead to generalization in at least some circumstances. In Experiment 3, there was another effect of English Frequency, but when the same talker produced each token of the different familiarization words, children were no more accurate to produce the clusters after hearing three words compared to hearing just one. It appears, then, that token variability and type frequency alone did not lead to phonotactic learning. Rather, it was the combination of the two that was critical for children to be able to learn phonotactics.

The results are most in favor of the model proposed by Pierrehumbert (2003b). This is because Pierrehumbert's model is the only one that correctly predicted that the combination of token variability and type frequency would generalize, but either factor in isolation would not. The exemplar model (Bailey & Hahn, 2001) is hampered by an unattested prediction that token variability alone would generalize, while the Albright (2007) model suffers from having incorrectly predicted that type frequency alone would generalize.

A better understanding of the results was sought by running additional analyses on the combined results of all four experiments. In the first analysis, the preponderance of perfect or near-perfect accuracy scores for the HiEng clusters led to the conclusion that the results are colored by a ceiling effect. This suggests that some effects of the Experimental Frequency manipulation may have been obscured, and additional experiments should be conducted to determine whether there were effects of Experimental Frequency, particularly in Experiments 1 and 3, that were hid by the ceiling effect. In the second analysis, a multiple regression showed that improved accuracy resulting from repeated repetitions, a measure of articultory practice, was less able to explain children's accuracy scores than the Experimental Frequency factor. Finally, the results were considered with respect to differences in syllable structure between the HiEng and LoEng items. Because of the ceiling effect, it is unlikely that syllable structure is responsible for the interaction of English Frequency and Experimental Frequency in Experiments 2A and 2B. However, syllable structure and phonotactic probability can not be separated in these experiments, and it is likely that the learning effects that were found are relevant to both factors.

8.3 Conclusion

The findings presented in this dissertation offer answers to the two questions that first motivated the experiments. With respect to the efficacy of perceptual learning in speech development, the effect Experimental Frequency in Experiments 2A and 2B on low probability clusters showed a clear role for perceptual learning. Furthermore, the results of a linear regression showed its role may be more substantial than the benefits gained from repeated articulations. With respect to question of how word-tokens and word-types contribute to phonotactic learning, the results showed that both were necessary for children to be able to generalize the target phonotactic sequences and that either factor on its own was not effective.

The consequences of these results can be extend past the particulars of the experiments. By showing how perceptual learning is relevant to the task of speech perception, these results have bridged a gap between the infant and child literatures on phonotactics. The results show for the first time how one learning mechanism, based in perception, can be responsible for both infant knowledge of phonotactics and phonotactic effects in child speech. Furthermore, the results challenge the claim that perceptual learning plays no role in the development of speech (Messum, 2007). The results also distinguish between three models of phonotactic learning and the representations that those models assume. The results show that the predictions of one model (Pierrehumbert's hybrid model) are more accurate than the predictions of the others. This turns out to be more than support for the model, however. It also supports an assumption of Pierrehumbert's model and of most modern linguistic research, that people have knowledge of abstract linguistic representations. The results provide convincing evidence that words and phonotactic sequences are psychologically real in the minds of speakers. The upshot is that we have convincing evidence for a phonological grammar.

Linguists reading this dissertation may find this conclusion to be somewhat anticlimactic. After all, many linguists assume that features and phonemes and words and syllables are real. However, support for this point of view has been hard to come by in the psychological and language acquisition literatures, so psychological evidence for linguistic representations should be taken as a boon for the assumptions of 100+ years of linguistic research.

By concluding that abstract forms are both real and learned, the conclusions reframe some of the debate about what we study when we study language acquisition. Exemplar models, which have become increasingly influential in many research programs, start by asking, "Given a set of minimally abstract units, what kinds of structures are simply reflexes of perception, storage, and use?" In contrast, the claim that abstractions are both real and learned is a return to the question, "How do children arrive at a symbolic adult grammar?" This is a partial return to a position advocated by Noam Chomsky (Chomsky, 1965) within the generative grammar framework, but it also has very important differences. "How" used to mean, "Look for evidence of linguistic representations in what infants/children say or do." I take "how" to mean, "How do children *learn* linguistic representations?" The important difference is that the old question assumed the prior existence of linguistic representations, but the new question does not.

An important component of the research program I am advocating is that it is still guided by many of the facts that first led people to assume that linguistic representations were innate. For example, research programs studying how children learn linguistic representations should be guided by two important facts: first, that learners tend to converge on similar grammars, and second, that grammars tend to employ similar units. These facts suggest that a bottom-up learner, such as the one I am advocating here, must be constrained in many ways, both in terms of what the learner is pre-equipped with, as well as what other aspects of speech perception and production can be learned from. In many ways, these are the same questions that have been asked for years by researchers in linguistics, psychology, and the speech and hearing sciences. That their explanation has defied explanation gives each field vitality and a common cause. That the present research provides evidence for just one among many proposals for answering these questions makes it an important contribution to anyone interested in how the fields might come together.

APPENDIX A

ITEM LISTS

Table A.1: The four item lists used in Experiment 1. The letters in the list title mark the assignment of clusters to blocks. The clusters kt, mp, ʃp, and fp were heard in Block 1 in Lists 1A and 2A, and in Block 2 for Lists 1B and 2B. The clusters sp, st, mk, and pk were heard in Block 1 for Lists 1B and 2B, and in Block 2 in Lists 1A and 2A.

	ENG FREQ	**ITEM LIST 1A**		**ITEM LIST 2B**		**EXP FREQ**
Block One	**HiEng**	**mp** fæmpɪm	**kt** boktəm	**sp** fospəm	**st** mæstəm	*ExpHi*
	LoEng	**ʃp** foʃpəm	**fp** mæfpəm	**mk** nʌmkəs	**pk** saʊpkəs	*ExpLo*
Block Two	**HiEng**	**sp** fospəm	**st** mæstəm	**mp** nʌmpəs	**kt** saʊktəs	*ExpLo*
	LoEng	**mk** fæmkɪm	**pk** bopkəm	**ʃp** foʃpəm	**fp** mæfpəm	*ExpHi*

	ENG FREQ	**ITEM LIST 1B**		**ITEM LIST 2A**		**EXP FREQ**
Block One	**HiEng**	**sp** fospəm	**st** mæstəm	**mp** nʌmpəs	**kt** saʊktəs	*ExpLo*
	LoEng	**mk** fæmkɪm	**pk** bopkəm	**ʃp** foʃpəm	**fp** mæfpəm	*ExpHi*
Block Two	**HiEng**	**mp** fæmpɪm	**kt** boktəm	**sp** fospəm	**st** mæstəm	*ExpHi*
	LoEng	**ʃp** foʃpəm	**fp** mæfpəm	**mk** nʌmkəs	**pk** saʊpkəs	*ExpLo*

Table A.2: The four familiarization item lists used in Experiment 2A. The letters in the list title mark the assignment of clusters to blocks. The clusters kt, mp, ʃp, and fp were heard in Block 1 in Lists 1A and 2A, and in Block 2 for Lists 1B and 2B. The clusters sp, st, mk, and pk were heard in Block 1 for Lists 1B and 2B, and in Block 2 in Lists 1A and 2A.

	ENG FREQ	**ITEM LIST 1A**		**ITEM LIST 2B**		**EXP FREQ**
Block One	**HiEng**	**mp**	**kt**	**sp**	**st**	*ExpHi*
		dompət	lɛktəf	kɛspəs	sʌstəp	
		sæmpəs	maʊktəs	tuspən	lostən	
		gumpət	saktəf	dɪspək	gɪstək	
	LoEng	**ʃp**	**fp**	**mk**	**pk**	*ExpLo*
		foʃpəm	mæfpəm	fæmkɪm	bopkəm	
Block Two	**HiEng**	**sp**	**st**	**mp**	**kt**	*ExpLo*
		fospəm	mæstəm	fæmpɪm	boktəm	
	LoEng	**mk**	**pk**	**ʃp**	**fp**	*ExpHi*
		domkət	lɛpkəf	kɛʃpəs	sʌfpət	
		sæmkəs	maʊpkəs	tuʃpən	lofpən	
		gumkət	sapkəf	dɪʃpək	gɪfpək	

	ENG FREQ	**ITEM LIST 1B**		**ITEM LIST 2A**		**EXP FREQ**
Block One	**HiEng**	**sp**	**st**	**mp**	**kt**	*ExpLo*
		fospəm	mæstəm	fæmpɪm	boktəm	
	LoEng	**mk**	**pk**	**ʃp**	**fp**	*ExpHi*
		domkət	lɛpkəf	kɛʃpəs	sʌfpət	
		sæmkəs	maʊpkəs	tuʃpən	lofpən	
		gumkət	sapkəf	dɪʃpək	gɪfpək	
Block Two	**HiEng**	**mp**	**kt**	**sp**	**st**	*ExpHi*
		dompət	lɛktəf	kɛspəs	sʌstəp	
		sæmpəs	maʊktəs	tuspən	lostən	
		gumpət	saktəf	dɪspək	gɪstək	
	LoEng	**ʃp**	**fp**	**mk**	**pk**	*ExpLo*
		foʃpəm	mæfpəm	fæmkɪm	bopkəm	

Table A.3: The four lists used in Experiments 2B and 3. Experiments 2B and 3 used identical word lists; on the association of word tokens and talkers was different. In Experiment 2B a different talker was assigned to each of the four word tokens; in Experiment 3 only one talker was assigned to the four word tokens. The letters in the list title mark the assignment of clusters to blocks. The clusters kt, mp, ʃp, and fp were heard in Block 1 in Lists 1A and 2A, and in Block 2 for Lists 1B and 2B. The clusters sp, st, mk, and pk were heard in Block 1 for Lists 1B and 2B, and in Block 2 in Lists 1A and 2A.

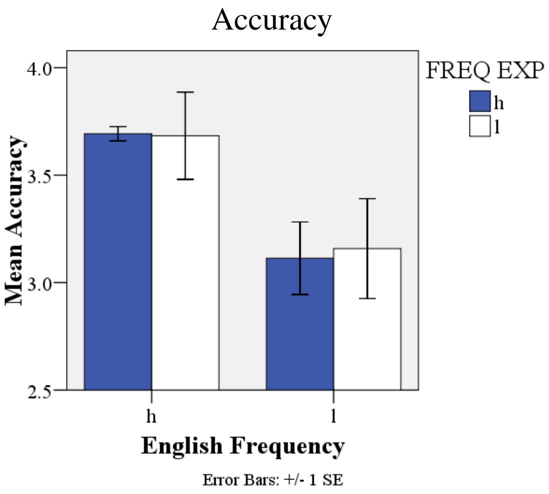
	ENG FREQ	**ITEM LIST 1A**		**ITEM LIST 2B**		**EXP FREQ**
Block One	**HiEng**	**mp**	**kt**	**sp**	**st**	*ExpHi*
		dɪmpət	lɛktəf	kɛspəs	mæstəm	
		nʌmpəs	saʊktəs	tuspən	baɪstəm	
		ɡumpən	biktəm	fospəm	ɡɪstək	
	LoEng	**ʃp**	**fp**	**mk**	**pk**	*ExpLo*
		foʃpəm	mæfpəm	nʌmkəs	saʊpkəs	
Block Two	**HiEng**	**sp**	**st**	**mp**	**kt**	*ExpLo*
		fospəm	mæstəm	nʌmpəs	saʊktəs	
	LoEng	**mk**	**pk**	**ʃp**	**fp**	*ExpHi*
		dimkət	lɛpkəf	kɛʃpəs	mæfpəm	
		nʌmkəs	saʊpkəs	tuʃpən	baɪfpəm	
		ɡumkən	bipkəm	foʃpəm	ɡɪfpək	

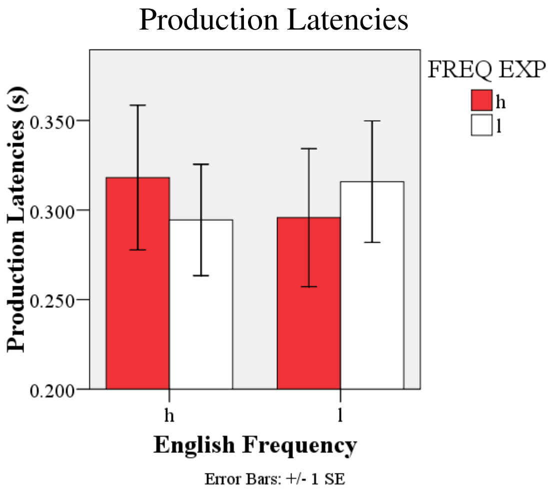
	ENG FREQ	**ITEM LIST 1B**		**ITEM LIST 2A**		**EXP FREQ**
Block Two	**HiEng**	**sp**	**st**	**mp**	**kt**	*ExpLo*
		fospəm	mæstəm	nʌmpəs	saʊktəs	
	LoEng	**mk**	**pk**	**ʃp**	**fp**	*ExpHi*
		dimkət	lɛpkəf	kɛʃpəs	mæfpəm	
		nʌmkəs	saʊpkəs	tuʃpən	baɪfpəm	
		ɡumkən	bipkəm	foʃpəm	ɡɪfpək	
Block One	**HiEng**	**mp**	**kt**	**sp**	**st**	*ExpHi*
		dɪmpət	lɛktəf	kɛspəs	mæstəm	
		nʌmpəs	saʊktəs	tuspən	baɪstəm	
		ɡumpən	biktəm	fospəm	ɡɪstək	
	LoEng	**ʃp**	**fp**	**mk**	**pk**	*ExpLo*
		foʃpəm	mæfpəm	nʌmkəs	saʊpkəs	

APPENDIX B

BY-ITEMS ANALYSES

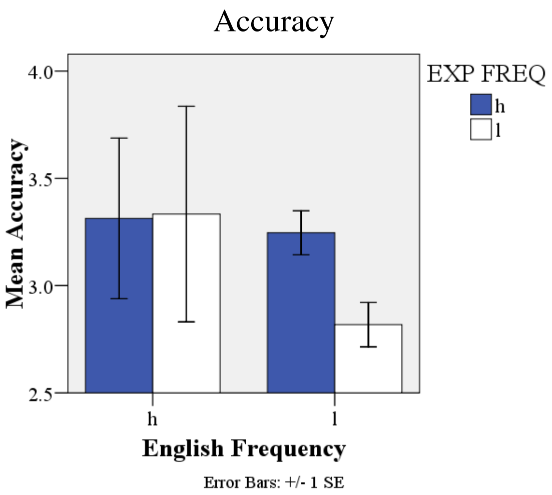
B.1 By-Items ANOVA - Experiment 1

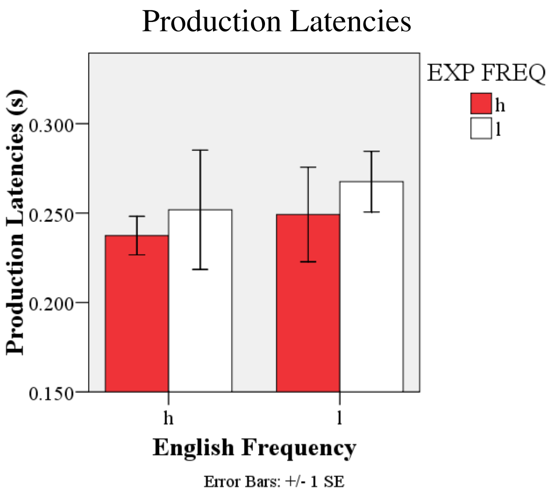




A mixed-design ANOVA was conducted for the items analysis, with English Frequency as a between-items factor and Experimental Frequency as a within-items factor. For the analysis of the accuracy data, there was a near-significant effect of English Frequency (F (1,6) = 5.895, p = .051, η_p^2 = .496), but no effect of Experimental Frequency and no interaction (both Fs $<$ 1). For the production latency analysis, there were no significant effects (all Fs $<$ 1). The results, while generally not significant, appear to reflect the same general pattern seen in the by-subjects analysis in Section 3.1.5 of Chapter 3, although all of the by-items analyses should be given limited interpretations due to the problems with the by-items analysis discussed in Section 3.1.2.

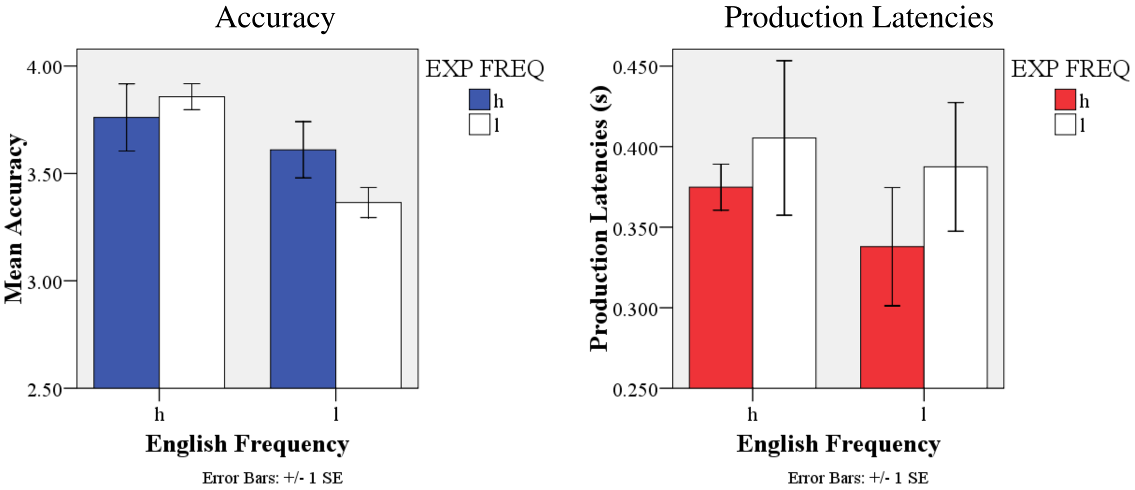
B.2 By-Items ANOVA - Experiment 2A





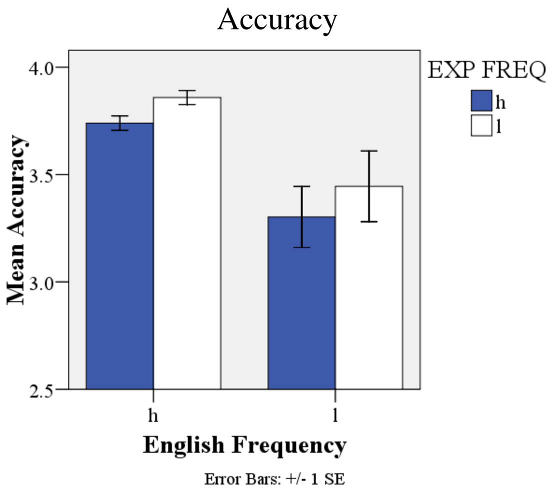
A mixed-design ANOVA was conducted for the items analysis, with English Frequency as a between-items factor and Experimental Frequency as a within-items factor. For the analysis of the accuracy data, there was not a significant effect of English Frequency ($F < 1$), but there was a significant effect of Experimental Frequency (F (1,5) = 7.617, $p < .05$, η_p^2 = .605) and a significant interaction (F (1,5) = 9.176, $p < .05$, η_p^2 = .647). The significant effect of Experimental Frequency and the interaction appear to be driven by the the very large effect of Experimental Frequency for the LoEng clusters. For the production latency analysis, there were no significant effects (all Fs < 2). The results generally reflect the same pattern seen in the by-subjects analysis in Section 4.1.5 of Chapter 4. Note that the degrees of freedom on this analysis are different than the other by-items analyses because one item, /zaspən/, was removed.

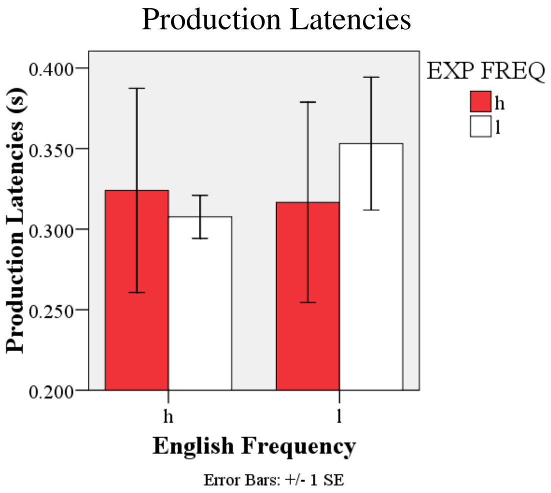
B.3 By-Items ANOVA - Experiment 2B



A mixed-design ANOVA was conducted for the items analysis, with English Frequency as a between-items factor and Experimental Frequency as a within-items factor. For the analysis of the accuracy data, there was a significant effect of English Frequency (F (1,6) = 8.857, $p < .05$, η_p^2 = .596), but not of Experimental Frequency and no significant interaction (both Fs < 2). For the production latency analysis, there was no significant effect of English Frequency ($F < 1$), but a near-significant effect of Experimental Frequency (F (1,6) = 4.731, p = .073, η_p^2 = .441) and no interaction ($F < 1$). These results appear to reflect the pattern of results found in the by-subjects analysis in Section 5.1.5 of Chapter 5.

B.4 By-Items ANOVA - Experiment 3





A mixed-design ANOVA was conducted for the items analysis, with English Frequency as a between-items factor and Experimental Frequency as a within-items factor. For the analysis of the accuracy data, there was a significant effect of English Frequency (F (1,6) = 8.032, $p < .05$, $\eta_p^2 = .572$), as well as a significant effect of Experimental Frequency (F (1,6) = 7.838, $p < .05$, $\eta_p^2 = .566$), but no interaction ($F < 1$). The effect of English Frequency appears to have resulted from less accurate productions when children were exposed to three words (ExpHi, blue bars) compared to when they were exposed to one word (ExpLo, white bars). This result is surprising given the non-significant by-subjects analysis, although both analyses seem to follow the same general trend. The results here are not given much credence, however, due to the problems discussed in Section 3.1.2. For the production latency analysis, there were no significant effects (all Fs $<$ 1). The results of the production latency analysis match the pattern of results foung in the by-subjects analysis in Section 6.1.5 of Chapter 6.

APPENDIX C

EXPERIMENTS 2B AND 3: PHONE AND BIPHONE SCORES

This appendix presents tables of the phone and biphone probabilities for the three item sets (ExpHi, ExpLo, and test items) used in Experiments 2B and 3. Due to the size of these tables, they are presented on separate pages.

Table C.1: Biphone probabilities for the ExpHi, ExpLo, and test item sets used in Experiment 2A. For the ExpHi set, which contains multiple words per cluster, the average biphone probability is given. Below these scores are the results of ANOVAs for the English Frequency and Set factors. Significant differences are given in bold text.

ENG FREQ	*CC*	BIPHONE1			BIPHONE2			BIPHONE3		
		ExpHi	*ExpLo*	*Test*	*ExpHi*	*ExpLo*	*Test*	*ExpHi*	*ExpLo*	*Test*
High	**kt**	0.0023	0.0007	0.0021	0.0031	0.0	0.001	0.0036	0.0036	0.0036
	mp	0.0057	0.0014	0.0046	0.0047	0.0051	0.0049	0.0091	0.0091	0.0091
	sp	0.0031	0.0062	0.0023	0.0036	0.0024	0.0024	0.0081	0.0081	0.0081
	st	0.0045	0.0101	0.0017	0.0084	0.0071	0.003	0.0232	0.0232	0.0232
Low	**fp**	0.0045	0.0101	0.0017	0.002	0.0013	0.0004	0.0	0.0	0.0
	mk	0.0057	0.0014	0.0046	0.0047	0.0051	0.0049	0.0002	0.0002	0.0002
	pk	0.0023	0.0007	0.0021	0.0015	0.0	0.0018	0.0002	0.0002	0.0002
	ʃp	0.0031	0.0062	0.0023	0.0003	0.0008	0.0002	0.0	0.0	0.0
	ENG FREQ	*F* (1,18) = .000, *p* = 1.00			*F* (1,18) = 4.115, *p* = .058			***F* (1,18) = 19.833, *p* < .001**		
	SET	*F* (2,18) = .977, *p* = .395			*F* (2,18) = .577, *p* = .572			*F* (2,18) = .000, *p* = 1.00		
	INTERACTION	*F* (2,18) = .000, *p* = 1.00			*F* (2,18) = .320, *p* = .730			*F* (2,18) = .000, *p* = 1.00		

		BIPHONE4			BIPHONE5			BIPHONE_SUM		
		ExpHi	*ExpLo*	*Test*	*ExpHi*	*ExpLo*	*Test*	*ExpHi*	*ExpLo*	*Test*
High	**kt**	0.0057	0.0057	0.0057	0.0076	0.0081	0.0169	0.0223	0.0181	0.0293
	mp	0.0042	0.0042	0.0042	0.0111	0.0081	0.0029	0.0348	0.0279	0.0257
	sp	0.0042	0.0042	0.0042	0.0122	0.0117	0.0029	0.0312	0.0325	0.0199
	st	0.0057	0.0057	0.0057	0.0088	0.0117	0.0169	0.0506	0.0577	0.0505
Low	**fp**	0.0042	0.0042	0.0042	0.0088	0.0117	0.0169	0.0195	0.0272	0.0232
	mk	0.0019	0.0019	0.0019	0.0111	0.0081	0.0029	0.0236	0.0167	0.0145
	pk	0.0019	0.0019	0.0019	0.0076	0.0081	0.0169	0.0135	0.011	0.0229
	ʃp	0.0042	0.0042	0.0042	0.0122	0.0117	0.0029	0.0199	0.0229	0.0095
	ENG FREQ	***F* (1,18) = 17.236, *p* < .01**			*F* (1,18) = .773, *p* = .474			***F* (1,18) = 10.836, *p* < .01**		
	SET	*F* (2,18) = .000, *p* = 1.00			*F* (2,18) = .777, *p* = .474			*F* (2,18) = .128, *p* = .880		
	INTERACTION	*F* (2,18) = .000, *p* = 1.00			*F* (2,18) = .773, *p* = .476			*F* (2,18) = .013, *p* = .987		

Table C.2: The phone probabilities for the ExpHi and ExpLo familiarization item sets and the test item set used in Experiment 2B. For the ExpHi set, which contains multiple words per cluster, the average phone probability is given. Below these phone scores are the results of ANOVAs for the English Frequency and Set factors. Significant differences between the sets are given in bold text.

ENG FREQ	*CC*	PHONE1			PHONE2			PHONE3		
		ExpHi	*ExpLo*	*Test*	*ExpHi*	*ExpLo*	*Test*	*ExpHi*	*ExpLo*	*Test*
High	**kt**	0.0626	0.1024	0.0445	0.0381	0.0097	0.0221	0.0535	0.0535	0.0535
	mp	0.0339	0.0238	0.1024	0.0525	0.0392	0.0794	0.0494	0.0494	0.0494
	sp	0.0613	0.0466	0.0518	0.0481	0.0493	0.0605	0.0788	0.0788	0.0788
	st	0.0448	0.0572	0.0238	0.07	0.0794	0.0292	0.0788	0.0788	0.0788
Low	**fp**	0.0448	0.0572	0.0238	0.07	0.0794	0.0292	0.0197	0.0197	0.0197
	mk	0.0339	0.0238	0.1024	0.0525	0.0392	0.0794	0.0494	0.0494	0.0494
	pk	0.0626	0.1024	0.0445	0.0381	0.0097	0.0221	0.0371	0.0371	0.0371
	ʃp	0.0613	0.0466	0.0518	0.0481	0.0493	0.0605	0.0077	0.0077	0.0077
ENG FREQ		*F* (1,18) = .000, p = 1.00			*F* (1,18) = .000, p = 1.00			***F* (1,18) = 27.208, $p < .001$**		
SET		*F* (2,18) = .127, p = .882			*F* (2,18) = .211, p = .812			*F* (2,18) = .000, p = 1.00		
INTERACTION		*F* (2,18) = .000, p = 1.00			*F* (2,18) = .000, p = 1.00			*F* (2,18) = .000, p = 1.00		

		PHONE4			PHONE5			PHONE6			PHONE_SUM		
		ExpHi	*ExpLo*	*Test*	*ExpHi*	*ExpLo*	*Test*	*ExpHi*	*ExpLo*	*Test*	*ExpHi*	*ExpLo*	*Test*
High	**kt**	0.0894	0.0894	0.0894	0.0816	0.0816	0.0816	0.035	0.0554	0.0832	0.3601	0.3919	0.3743
	mp	0.0362	0.0362	0.0362	0.0816	0.0816	0.0816	0.0798	0.0554	0.014	0.3335	0.2857	0.3631
	sp	0.0362	0.0362	0.0362	0.0816	0.0816	0.0816	0.058	0.0355	0.044	0.364	0.328	0.353
	st	0.0894	0.0894	0.0894	0.0816	0.0816	0.0816	0.0383	0.0355	0.0832	0.4029	0.4219	0.3861
Low	**fp**	0.0362	0.0362	0.0362	0.0816	0.0816	0.0816	0.0383	0.0355	0.0832	0.2906	0.3096	0.2738
	mk	0.0422	0.0422	0.0422	0.0816	0.0816	0.0816	0.0798	0.0554	0.014	0.3395	0.2917	0.3692
	pk	0.0422	0.0422	0.0422	0.0816	0.0816	0.0816	0.035	0.0554	0.0832	0.2967	0.3285	0.3108
	ʃp	0.0362	0.0362	0.0362	0.0816	0.0816	0.0816	0.058	0.0355	0.044	0.2929	0.2569	0.2819
ENG FREQ		***F* (1,18) = 6.995, $p < .05$**			*F* (1,18) = .000, p = 1.00			*F* (1,18) = .000, p = 1.00			***F* (1,18) = 16.032, $p < .01$**		
SET		*F* (2,18) = .000, p = 1.00			*F* (2,18) = .000, p = 1.00			*F* (2,18) = .422, p = .662			*F* (2,18) = .230, p = .797		
INTERACTION		*F* (2,18) = .000, p = 1.00			*F* (2,18) = .000, p = 1.00			*F* (2,18) = .000, p = 1.00			*F* (2,18) = .000, p = 1.00		

REFERENCES

Albright, A. (2002). Islands of reliability for regular morphology: Evidence from italian. *Language*, *78*, 684-709.

Albright, A. (2007). *Gradient phonological acceptability as a grammatical effect.* obtained online from http://web.mit.edu/albright/www/papers/Albright-GrammaticalGradience.pdf.

Albright, A., & Hayes, B. (2002). Modeling english past tense intuitions with minimal generalization. In M. Maxwell (Ed.), *Proceedings of the 6th meeting of the acl special interest group in computational phonology.* New Brunswick, NJ: Association for Computational Linguistics. obtained online from http://web.mit.edu/albright/www/papers/AlbrightHayes02.pdf.

Albright, A., & Hayes, B. (2003). Rules vs. analogy in english past tenses: A computational/experimental study. *Cognition*, *90*, 119-161.

Association, I. P. (1999). *Handbook of the international phonetic association*. New York, NY: Cambridge University Press.

Baayen, R. H. (2004). Statistics in psycholinguistics: A critique of some current gold standards. In *Mental lexicon working papers i* (p. 1-45). University of Alberta, Edmonton. obtained from http://www.mpi.nl/world/persons/private/baayen/publications/Statistics.pdf.

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, *4*, 568-591.

Beckman, M. E., & Edwards, J. (1999). Lexical frequency effects on young children's imitative productions. In M. B. Broe & J. B. Pierrehumbert (Eds.), *Papers in laboratory phonology v: Acquisition and the lexicon* (p. 208-218). Cambridge, MA: Cambridge University Press.

Berent, I., Lennertz, T., Jun, J., Moreno, M. A., & Smolensky, P. (2008). Language universals in human brains. *Proceedings of the National Academy of Sciences*, *105(14)*, 5321-5325.

Berko, J. (1958). The child's learning of english morphology. *Word*, *14*, 150-177.

Bernhardt, B., & Stemberger, J. (1998). *Handbook of phonological development*. San Diego, CA: Academic Press.

Boysson-Bardies, B. d., Hallé, P., Sagart, L., & Durand, C. (1989). A crosslinguistic investigation of vowel formants in babbling. *Journal of Child Language*, *16(1)*, 1-17.

Boysson-Bardies, B. d., Sagart, L., & Durand, C. (1984). Discernible differences in the babbling of infants according to target language. *Journal of Child Language*, *11(1)*, 1-15.

Boysson-Bardies, B. d., & Vihman, M. M. (1991). Adaptation to language: Evidence from babbling and first words in four languages. *Language*, *67(2)*, 297-319.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training japanese listeners to identify english /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics*, *61(5)*, 977-985.

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

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, *10(5)*, 225-255.

Bybee, J. (2001). *Phonology and language use*. Boston, MA: Cambridge University Press.

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, *87*, B69-B77.

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.

Chin, S. B., & Dinnsen, D. A. (1992). Consonant clusters in disordered speech: Constraints and correspondence patterns. *Journal of Child Language*, *19(2)*, 259-285.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N., & Halle, M. (1968). *The sound pattern of english*. New York, NY: Harper and Row.

Clahsen, H., & Rothweiler, M. (1992). Inflectional rules in children's grammars: Evidence from the development of participles in german. *Yearbook of Morphology 1992*, 1-34.

Coleman, J., & Pierrehumbert, J. (1997, 12 July 1997). *Stochastic phonological grammars and acceptability* (Tech. Rep.). Association for Computational Linguistics, Somerset NJ. (obtained online from http://www.ling.northwestern.edu/ jbp/publications/publications.html, March 2008)

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers*, *28*, 125-127.

DeCasper, A. J., & Fifer, W. P. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science*, *208*, 1174-1176.

Edwards, J., & Beckman, M. E. (2008). Some cross-linguistic evidence for modulation of implicational universals by language-specific frequency effects in the acquisition of consonant phonemes. *Language Learning & Development*, *4(1)*, 122-156.

Edwards, J., Beckman, M. E., & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research*, *47(2)*, 421-436.

Elbert, M., & Gierut, J. A. (1986). *Handbook of clinical phonology: Approaches to assessment and treatment*. London: Taylor & Francis.

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior & Development*, *8*, 181-195.

Flege, J. E. (1993). Production and perception of a novel, second-language phonetic contrast. *Journal of the Acoustical Society of America*, *93*, 1589-1608.

Flege, J. E. (1995). Second-language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (p. pp. 229-273). Timonium, MD: York Press. obtained online from http://jimflege.com/bookchapters.html on 2/25/2008.

Flege, J. E. (1999). The relation between l2 perception and production. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (Eds.), *Proceedings of the xivth international congress of phonetic sciences* (p. 1273-1276). San Francisco, CA: obtained online from http://jimflege.com/conferenceproceedings.html on 2/25/2008.

Flege, J. E., Bohn, O.-S., & Jang, S. (1997). The production and perception of english vowels by native speakers of german, korean, mandarin, and spanish. *Journal of Phonetics*, *25*, 437-470.

Frauenfelder, U. H., & Schreuder, R. (1992). Constraining psycholinguistic models of morphological processing and representation: The role of productivity. In B. Booij & J. van Marle (Eds.), *Yearbook of morphology 1991.* The Netherlands: Kluwer.

Friederici, A. D., & Wessels, J. M. I. (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception & Psychophysics*, *54(3)*, 287-295.

Gathercole, S. E., & Martin, A. J. (1996). Interactive processes in phonological memory. In M. A. Conway (Ed.), *Cognitive models of memory.* Hove, UK: Psychology Press/MIT Press.

Gathercole, S. E., Willis, C., Emslie, H., & Baddeley, A. (1991). The influence of number of syllables and wordlikeness on children's repetition of nonwords. *Applied Psycholinguistics*, *12*, 349-367.

Gerken, L. A., Goffman, L., Carter, A., Bollt, A., Bruner, A., Fava, E., et al. (2006). *Statistical frequency in perception affects children's lexical production.* Unpublished Manuscript.

Gibbon, F. (1999). Undifferentiated lingual gestures in children with articulatory/phonological disorders. *Journal of Speech, Language, and Hearing Research*, *42*, 382-397.

Goffman, L., Gerken, L. A., & Lucchesi, J. (2007). Relations between segmental and motor variability in prosodically complex nonword sequences. *Journal of Speech, Language, and Hearing Research*, *50*, 444-458.

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

Goldinger, S. D. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological Review*, *105(2)*, 251-279.

Goldinger, S. D. (2007). A complementary-systems approach to abstract and episodic speech perception. In *International conference of phonetic sciences xvi.* obtained from http://www.ichps2007.de/conference/Papers/1781/1781.pdf.

Gómez, R. L., & Gerken, L. A. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*, 109-135.

Gómez, R. L., & Gerken, L. A. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, *4(5)*, 178-186.

Greenberg, J. H., & Jenkins, J. J. (1964). Studies in the psychological correlates of the sound system of american english. *Word*, *20*, 157-177.

Hammond, M. (2003). Phonotactics and probabilistic ranking. In A. Carnie, H. Harley, & M. Willie (Eds.), *Formal approaches to function in grammar: In honor of eloise jelinek* (p. 319-332). Amsterdam: John Benjamins.

Hammond, M. (2004). Gradience, phonotactics, and the lexicon in englislh phonology. *International Journal of English Studies*, *4(2)*.

Hay, J., Pierrehumbert, J. B., & Beckman, M. E. (2003). Speech perception, well-formedness and the statistics of the lexicon. In J. Local, R. Ogden, & R. Temple (Eds.), *Phonetic interpretation: Papers in laboratory phonology vi* (p. 58-74). Cambridge, MA: Cambridge University Press.

Hintzman, D. L. (1986). "schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93(4)*, 411-428.

Hodson, B. W., & Paden, E. P. (1981). Phonological processes which characterize unintelligible and intelligible speech patterns. *Journal of Speech and Hearing Disorders*, *46*, 369-373.

Hollich, G., Jusczyk, P. W., & Luce, P. A. (2002). Lexical neighborhood effects in 17-month-old word learning. In *Proceedings of the 26th annual boston university conference on language development* (p. 314-323). Boston, MA: Cascadilla Press.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *Journal of the Acoustical Society of America*, *119(5)*, 3059-3071.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology, Human Perception and Performance*, *26(5)*, 1570-1582.

Houston, D. M., & Jusczyk, P. W. (2003). Infants' long-term memory for the sound patterns of words and voices. *Journal of Experimental Psychology: Human Perception and Performance*, *29(6)*, 1143-1154.

Jespersen, O. (1904). *Lehrbuch der phonetik*. Leipzig and Berlin.

Johnson, K. (1997). Speech perception and speaker normalization. In J. K. & J. Mullennix (Eds.), *Talker variability in speech processing.* San Diego, CA: Academic Press.

Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

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

Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Preference for the predominant stress patterns of english words. *Child Development*, *64(3)*, 675-687.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M., Svenkerud, V. Y., & Jusczyk, M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, *32(3)*, 402-420.

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.

Kemler Nelson, D. G., Jusczyk, P. W., Mandel, D. R., Myers, J., Turk, A., & Gerken, L. A. (1995). The head-turn preference procedure for testing auditory perception. *Infant Behavior and Development*, *18(1)*, 111-116.

Kenstowicz, M. (1994). *Phonology in generative grammar.* Oxford, UK: Blackwell.

Kisseberth, C. W. (1970). On the functional unity of phonological rules. *Linguistic Inquiry*, *1*, 291-306.

Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, *12*, 119-131.

Locke, J. L. (1983). *Phonological acquisition and change.* New York, NY: Academic Press.

Logan, J. D., Lively, S. E., & Pisoni, D. B. (1991). Training japanese listeners to identify english /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, *89*, 2076-2087.

Lotto, A. J., Sato, M., & Diehl, R. L. (2004). Mapping the task for the second language learner: the case of japanese acquisition of /r/ and /l/. In *From sound to sense.* obtained from http://www.rle.mit.edu/soundtosense/conference/pdfs/fulltext/ Saturday

Luce, P. A. (1986). *Neighborhoods of words in the mental lexicon* (Tech. Rep.). Speech Research Laboratory, Department of Psychology: Indiana University.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*, 1-36.

Luce, P. A., Pisoni, D. B., & Goldinger, S. D. (1990). Similarity neighborhoods of spoken words. In G. T. M. Altman (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (p. 105-121). Cambridge, MA: MIT Press.

MacKay, D. G. (1989). *The organization of perception and action: A theory for language and other cognitive skills.* New York, NY: Springer-Verlag.

MacWhinney, B., & Leinbach, J. (1990). Implementations are not conceptualizations: Revising the verb learning model. *Cognition*, *29*, 121-157.

Maekawa, J., & Storkel, H. L. (2006). Individual differences in the influence of phonological characteristics on expressive vocabulary development by young children. *Journal of Child Language*, *33*, 439-459.

Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., Woest, A., & Pinker, S. (1993). *German inflection: The exception that proves the rule.* Occasional Paper No. 47. Center for Cognitive Science, MIT, Cambridge, MA.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., & Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, *57(4)*.

Marcus, G. F., Vijayan, S., Bandi-Rao, S., & Vishton, P. M. (1999). Do infants learn grammar with algebra or statistics? *Science*, *284*, 436-437.

Mattys, S. L., & Jusczyk, P. W. (2000). Phonotactic cues for segmentation of fluid speech by infants. *Cognition*, *78*, 91-121.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, *38*, 465-494.

Maye, J., & Gerken, L. A. (2000). Learning phonemes without minimal pairs. In *Proceedings of the 24th annual boston university conference on language development* (p. 522-533).

Maye, J., Werker, J. F., & Gerken, L. A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82(3)*, B101-B111.

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

McLeod, S., Doorn, J. van, & Reed, V. A. (2001). Normal acquisition of consonant clusters. *American Journal of Speech-Language Pathology*, *10*, 99-110.

Mehler, J., Bertoncini, J., Barrière, M., & Jassik-Gerschenfeld, D. (1978). Infant recognition of mother's voice. *Perception*, *7(5)*, 491-497.

Mehler, J., Dupoux, E., & Segui, J. (1990). Constraining models of lexical access: The onset of word recognition. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing* (p. 236-262). Hillsdale, NJ: Erlbaum.

Mehler, J., Jusczyk, P. W., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29(2)*, 143-178.

Messer, S. (1967). Implicit phonology in children. *Journal of Verbal Learning & Verbal Behavior*, *6(4)*, 609-613.

Messum, P. R. (2007). *The role of imitation in learning to pronounce*. Unpublished doctoral dissertation, London University, London, UK. (obtained online from http://p.messum.googlepages.com/downloads, April 2008)

Munson, B. (2001). Phonological pattern frequency and speech production in adults and children. *Journal of Speech, Language, and Hearing Research*, *44(4)*, 778-792.

Munson, B., Swenson, C. L., & Manthei, S. C. (2005). Lexical and phonological organization in children: Evidence from repetition tasks. *Journal of Speech, Language, and Hearing Research*, *48*, 108-124.

Newman, A. J., Ullman, M. T., Pancheva, R., Waligura, D. L., & Neville, H. J. (2007).

An erp study of regular and irregular english tense inflection. *NeuroImage*, *34*, 435-445.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*, 198-234.

Nosofsky, R. M. (1986). Identity, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115(1)*, 39-57.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14(4)*, 700-708.

Nosofsky, R. M., & Bergert, F. B. (2007). Limitations of exemplar models of multi-attribute probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 999-1019.

Nosofsky, R. M., & Zaki, S. R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. *Psychological Science*, *9(4)*, 247-255.

Ohala, D. K. (1996). *Cluster reduction and constraints on acquisition*. Unpublished doctoral dissertation, University of Arizona, Tucson, AZ.

Ohala, D. K. (1999). The influence of sonority on children's cluster reductions. *Journal of Communication Disorders*, *32(6)*, 397-422.

Ohala, J., & Ohala, M. (1986). Testing hypotheses regarding the psychological manifestation of morpheme structure constraints. In *Experimental phonology* (p. 239252). Orlando, FL: Academic Press.

Olmsted, D. L. (1971). *Out of the mouths of babes*. The Hague: Mouton.

Pertz, D. L., & Bever, T. G. (1975). Sensitivity to phonological universals in children and adolescents. *Language*, *51*, 149-162.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In P. Hopper & J. Bybee (Eds.), *Frequency and the emergence of linguistic structre* (p. 137-158). Philadelphia, PA: John Benjamins.

Pierrehumbert, J. B. (2003a). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech*, *46(2-3)*, 115-154.

Pierrehumbert, J. B. (2003b). Probabilisitic phonology: Discrimination and robustness. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics* (p. 177-228). Cambridge, MA: MIT Press.

Pierrehumbert, J. B. (2006). The next toolkit. *Journal of Phonetics*, *34(4)*, 516-530.

Pinker, S. (1991). Rules of language. *Science*, *253*, 530-535.

Pinker, S. (1999). *Words and rules*. New York, NY: HarperCollins Publishers.

Pinker, S., & Prince, A. (1994). Regular and irregular morphology and the psychological status of rules of grammar. In S. D. Lima, R. L. Corrigan, & G. K. Iverson (Eds.), *The reality of linguistic rules* (p. 353-388). Amsterdam: John Benjamins.

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6(11)*, 456-463.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition*, *38*, 43-102.

Prasada, S., & Pinker, S. (1993). Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes*, *8*, 1-56.

Prather, E. M., Hedrick, D. L., & Kern, C. A. (1975). Articulation development in children aged two to four years. *Journal of Speech and Hearing Disorders*, *40*, 179-191.

Prince, A., & Smolensky, P. (1993). *Optimality theory.* Rutgers University and University of Colorado.

Raaijmakers, J. G. W., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to deal with "the language-as-fixed-effect" fallacy: Common misconceptions and alternative solutions. *Journal of Memory and Language*, *41*, 416-426.

Reber, A. S. (1963). On learning the grammatical order of words. *Psychological Review*, *70*, 323-348.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *77*, 317-327.

Rumelhart, D., & McClelland, J. (1986). On earning the past tenses of english verbs: Implicit rules or parallel distributed processing? In J. McClelland, D. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations of the microstructure of cognition.* Cambridge, MA: MIT Press.

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

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

Schwartz, R. G., & Leonard, L. B. (1982). Do children pick and choose? an examination of phonological selection and avoidance in early lexical acquisition. *Journal of Child Language*, *9(2)*, 319-336.

Shriberg, L., & Kwiatowski, J. (1980). *Natural process analysis*. New York, NY: John Wiley.

Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition*, *106(2)*, 833-870.

Skousen, R. (1989). *Analogical modeling of language*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Smit, A. B. (1993). Phonologic error distributions in the iowa-nebraska norms project: Word-initial consonant clusters. *Journal of Speech and Hearing Research*, *36*, 931-947.

Smit, A. B., Hand, L., Freilinger, J. J., Bernthal, J. E., & Bird, A. (1990). The iowa articulation norms project and its nebraska replication. *Journal of Speech and Hearing Disorders*, *55*, 779-798.

Smith, J. D. (2005). Wanted: A new psychology of exemplars. *Canadian Journal of Psychology*, *59(1)*, 47-53.

Smith, J. D., & Minda, J. P. (2002). Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *28(4)*, 800-811.

Stevens, K. N. (1998). *Acoustic phonetics*. Cambridge, MA: MIT Press.

Storkel, H. L. (2001). Learning new words: Phonotactic probabilities in language development. *Journal of Speech, Language, and Hearing Research*, *44*, 1321-1337.

Storkel, H. L. (2003). Learning new words ii: Phonotactic probability in verb learning. *Journal of Speech, Language, and Hearing Research*, *46*, 1312-1323.

Storkel, H. L. (2004). The emerging lexicon of children with phonological delays. *Journal of Speech, Language, and Hearing Research*, *47*, 1194-1212.

Storkel, H. L., & Morrisette, M. L. (2002). The lexicon and phonology: Interactions in language acquisition. *Language, Speech, and Hearing Services in Schools*, *33*, 24-37.

Storkel, H. L., & Rogers, M. A. (2000). The effect of probabilistic phonotactics on lexical acquisition. *Clinical Linguistics & Phonetics*, *14*, 407-425.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science*, *9(4)*, 325-329.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*, 374-408.

Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in english. *Behavior Research Methods, Instruments, & Computers*, *36(3)*, 481-487. Obtained from www.people.ku.edu/ mvitevit/PhonoProbHome.html.

Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer, E. T. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language*, *68*, 306-311.

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of mandarin tone production before and after perceptual training. *Journal of the Acoustical Society of America*, *113(2)*, 1033-1043.

Werker, J. F., Gilbert, J. H. V., Humphrey, K., & Tees, R. C. (1984). Developmental aspects of cross-language speech perception. *Child Development*, *52*, 349-355.

Werker, J. F., & Tees, R. C. (2002). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *25(1)*, 121-133.

Woodward, J. Z., & Aslin, R. N. (1990). *Segmentation cues in maternal speech to infants.* Paper presetned at the 7th bienneial meeting of the International Conference on Infant Studies, Montreal, Quebec, Canada.

Zamuner, T. S., Gerken, L. A., & Hammond, M. (2004). Phonotactic probabilities in young children's speech production. *Journal of Child Language*, *31(3)*, 515-536.