Carnegie Mellon University
Research Showcase @ CMU

Dietrich College Honors Theses | Dietrich College of Humanities and Social Sciences

4-23-2009

Segmenting the Signal: A Cross-Age Comparison of Speech and Tone Stimuli

Alexandra T. Kronstein
Carnegie Mellon University, atk@cmu.edu

Follow this and additional works at: http://repository.cmu.edu/hsshonors

This Thesis is brought to you for free and open access by the Dietrich College of Humanities and Social Sciences at Research Showcase @ CMU. It has been accepted for inclusion in Dietrich College Honors Theses by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.

Running head: A CROSS-AGE COMPARISON OF SPEECH AND TONE STIMULI

Segmenting the Signal: A Cross-Age Comparison of Speech and Tone Stimuli

Alexandra T. Kronstein

Carnegie Mellon University

Advisor: Erik D. Thiessen, Ph.D.

Address for correspondence:

Alexandra T. Kronstein

Department of Psychology

Carnegie Mellon University

5000 Forbes Avenue

Pittsburgh, PA 15213

atk@cmu.edu

Abstract

The present study sought to extend findings of an "embeddedness constraint" on statistical learning in the visual domain (Fiser & Aslin, 2005) to the auditory domain. In this set of experiments, we explored the possible manifestations of an embeddedness constraint on speech and tone sequences in studies with adult and infant participants. An embeddedness constraint operant in the linguistic domain could benefit learners by helping them to circumvent the "combinatorial explosion" that arises from large amounts of complex data. This constraint could also prove useful in language acquisition tasks, such as word segmentation, allowing learners to advance more quickly to higher-order tasks, such as semantic and syntactic categorization. Furthermore, findings of an embeddedness constraint on both speech and tone stimuli might suggest that this constraint is broadly domain-general. Present results do not indicate an embeddedness constraint on speech or tone input in adults. However, pending studies with infants may suggest that the existence of the embeddedness constraint is itself constrained by maturation or by experience with language.

Segmenting the Signal: A Cross-Age Comparison of Speech and Tone Stimuli

One of the most challenging questions researchers must answer is how children acquire language so easily when the input is so complex. Comprehension of any given component of language entails a detailed knowledge of many dimensions. For example, to understand one single word, an infant must be familiar with the speech sounds of her native language, how these sounds generalize across speakers, and the word's boundaries when the target item is spoken in fluent speech. Despite such intricacy, children elegantly synthesize the disparate parts of language into a functioning whole with greater ease and aplomb than adults (Newport, 1990). However, the complexity of language is not the only problem young learners face when acquiring their first language or languages. Infants must also learn to map between different complex aspects of the input. For any given word, there are an infinite number of possible meanings (Quine, 1960). For example, a child may see a big, red, bouncy ball, and hears the label "ball!" Yet there is no reason for that child to assume that "ball" refers to the object as a whole. It could just as easily refer to a quality of the ball, such as "red" or "bouncy." Even in such a simple interaction the possibilities are endless. This complexity inherent in language, the seemingly infinite possibilities from the input, results in a formidable challenge known as the "curse of dimensionality" or the "combinatorial explosion." As it relates to language, the curse of dimensionality describes the difficulty of identifying the correct set of rules from infinite possibilities given only a limited number of exemplars.

In an attempt to explain how learners circumvent this problem, researchers have amassed a large corpus of evidence to suggest that for both infants and adults to learn, there must be constraints on either the mechanism by which they learn or constraints on the input they receive (Yang, 2004; Pinker, 1994; Thiessen, 2009; Saffran, et al, 1996b). Although the idea of

constraints is almost universally accepted by researchers, the nature of these constraints and the definition of the mechanism are sources of great disagreement. Some theorists regard both the mechanism and constraints on learning as domain-specific, or unique to linguistic stimuli, and possibly as biologically or genetically unique to humans. Others view both the mechanism of and constraints on learning as domain-general, or applicable to many different types of input, and possibly present across species. Broadly speaking, the two primary perspectives on language acquisition, Universal Grammar and Learning Theory, have arisen as a result of differing predictions about the nature of constraints and the type of mechanism that governs language acquisition. Nativists view constraints on language acquisition and the mechanism as domain-specific, whereas learning theorists see constraints on language acquisition and the mechanism as domain-general. In order to understand the sometimes subtle differences between predictions made by advocates of either theory, knowledge of the major developments of both perspectives is necessary.

The perspective that the mechanism is a uniquely human endowment with language-specific characteristics stems from Chomsky's (1959) unequivocal rejection of behaviorist principles as outlined in *A Review of B.F. Skinner's Verbal Behavior*. Most famously instantiated in the "Universal Grammar" theory introduced by Chomsky, the nativist claim rests on the idea that language is too complex of a system to be learned without preexisting, innate knowledge of its general concepts (Smith, 2005). In its early years, the nativist position argued that a Language Acquisition Device (LAD) governed primary language acquisition (Chomsky, 1965), but more recently this theory has evolved into a set of universal language principles with language-specific parameters (Culicover, 1997). The principles and parameters approach to language acquisition relies on a set of universal principles found only in human languages.

According to this account, language-specific parameters are activated or deactivated depending on which language (or languages) a child is exposed to. The nativist position has found much support from linguists who advocate the "unity in variety" (Kager, 1999) approach very elegantly instantiated in linguistic models such as Optimality Theory (Prince & Smolensky, 2004). Optimality Theory resolves seemingly random nuances in the languages of the world by ranking a set of constraints on language. This model of universal grammar allows for an unlimited set of candidates from which the "optimal" candidate is selected by ranking and evaluating different language-universal constraints. The optimal candidate is the one which violates the fewest high-ranking constraints (see Kager, 1999 for discussion and examples).

Central to all nativist theories of language acquisition is the idea that the input alone does not contain enough information for a child to acquire language. Based on Gold's (1967) proof, which asserts that a language cannot be learned by positive evidence alone, Chomsky's (1980) "poverty of the stimulus" argument forms the foundation for the "logical problem of language acquisition." The logical problem of language acquisition arises from a lack of negative evidence. Negative evidence is provided to the learner through corrections to the learner's output. This is in contrast to positive evidence, which consists of the set of syntactically correct examples a learner hears in the input. The problem with learning from positive evidence alone (and thereby the alleged need for negative evidence) is that positive evidence can be consistent with multiple different hypotheses about the underlying structure of the input. For example, without negative evidence, a Francophone child could hypothesize that all adjectives follow the nouns which they modify. However, this hypothesis, while supported by much positive evidence, is not correct. Adjectives describing beauty, age, goodness, or size actually precede the nouns which they modify. The nativist solution to this logical problem is a set of innate,

domain-specific constraints that assist the learner in acquiring his native language. These language-specific constraints are often referred to as a "universal grammar."

It is important to note that Universal Grammar refers to both the domain-specific mechanism that enables children to acquire language as well as an actual manifestation of "universal grammar" that is common to all human languages. This universal grammar explores the abstract similarities between human languages (e.g. CV is the universal unmarked syllabic structure (Zamuner, 2003)) and explains them in terms of the biological or genetic mechanism of Universal Grammar (see Comrie, 1989 for a detailed discussion). The mechanism of Universal Grammar is defined as a species-specific, domain-specific, language module that humans are born with but then decays or disappears with time (Chomsky, 1965). This biological or genetic mechanism is thought to constrain the preferred types of cognitive representations that subsequently appear in the data.

Both the mechanism and the manifestation of universal grammar can be viewed as constraints on language learning. Indeed, the manifestation of Universal Grammar operates as a constraint on the input (Bickerton, 1981), and both the mechanism and the manifestation constrain the output of the human (Landau, 1998). However, the interaction between the mechanism of Universal Grammar and its manifestation poses a problem for those researchers interested in language evolution. Christiansen and Chater (2008) argue that an insurmountable flaw in the Universal Grammar theory of language acquisition is the "*circularity trap* (5)" because it poses a classic chicken-and-egg problem: which came first, the mechanism or the manifestation? As the mechanism of Universal Grammar is purported to be highly specific, it seems unlikely that all aspects present in the mechanism could have (or would have) evolved suddenly through one single chance mutation. If the mechanism of Universal Grammar evolved

over time, it may have been shaped by demands independent of the original mechanism, such as production limitations or social pressures. Such an explanation is inconsistent with a central claim of Universal Grammar: that the input is constrained by the mechanism and not vice versa.

Conversely, a Learning Theory of language acquisition allows for interaction between the input and the mechanism. Rather than viewing language acquisition as the result of a uniquely human mechanism with domain-specific parameters, advocates of learning theory assert that language is simply the result of domain-general learning at which humans excel. Learning theorists argue that powerful domain-general mechanisms, such as statistical learning, are capable of apprehending complex patterns across the sensory modalities. Statistical learning refers to the ability to derive patterns from the input based on elements of the input that correlate with or predict each other (Thiessen, 2009). The term "statistical learning" implies that the learner computes some type of implicit statistical analysis on the input, although the input itself does not need to be "statistical" or regular in nature. However, without regular input, the output of statistical learning cannot be particularly useful.

Unlike the domain-specific mechanism proposed by nativists, learning theorists assert that a domain-general mechanism governs language acquisition. This domain-general mechanism should be present across modalities, ages, and perhaps even across species, and should be useful in many different tasks. Statistical learning may be a representation of this domain-general mechanism, as it is present in both infant (Saffran, et al., 1996a) and adult humans (Saffran, et al., 1996b), as well as in non-human primates (Hauser, et al., 2001). Evident across modalities (Conway & Christiansen, 2006), it has been demonstrated that statistical learning is useful in many diverse tasks, from extracting visual feature hierarchies from multi-element scenes (Fiser & Aslin, 2005) to learning an artificial grammar from sequences of tactile

stimuli (Conway & Christiansen, 2005). In spite of the wealth of data on the pervasive presence of statistical learning, little is known about the manner in which the statistical analysis is conducted. However, the unifying point across such diverse studies is that learners are able to detect the structural patterns of the input without innate hypotheses of what that structure may be. One problem facing researchers across modalities is that the structure of the input can be thought of in several different ways. For example, in language, the structure can be viewed as associative in nature (i.e. individual units are combined bottom-up based on the probabilities between the units), as Bayesian in nature (i.e. every unit is subject to Bayesian hypothesis testing), or as chunks (i.e. chunks are selected with respect to working memory and processing demands and then undergo statistical analysis).

The regularity of the input is one of many reasons advocates of Learning Theory view the input as a rich source of data, unlike proponents of Universal Grammar. Although Gold's (1967) proof claims that a language cannot be learned by positive evidence alone, several researchers have demonstrated both infants' and adults' ability to use only positive evidence from the input to learn a number of tasks relevant for language acquisition. Using only positive evidence from highly controlled input, statistical learning has helped infants to track the distribution of phones in their native language (Maye, et al., 2002), to distinguish between known and unknown words in fluent speech (Jusczyk & Aslin, 1995), and to acquire the past-tense form of irregular English verbs (Kielar, et al., 2008). Thus, the nativist logical problem of language acquisition, which arises from an impoverished stimulus, may be solved simply by considering the richness of the base. Although researchers have investigated the ability of statistical learning to derive elegant patterns from only positive information in the input both within and across different domains (Conway & Christiansen, 2005), it is statistical learning's applicability in the linguistic domain

that is most contested by advocates of nativist theories of language acquisition. Nativists do not dispute that a statistical learning mechanism exists in the visual domain and can facilitate the learning of patterns of non-linguistic shapes in both infants (Kirkham et al., 2002) and adults (Fiser & Aslin, 2001). Nor do proponents of Universal Grammar challenge the finding that both infants and adults can use constrained statistical learning to extract relevant tone sequences from a string of tones (Saffran et al., 1999) as this is non-linguistic auditory input.

Both supporters of Universal Grammar and supporters of Learning Theory agree that humans have both innate and learned capabilities. For example, it has been suggested that humans possess an innate cognitive mechanism for predator recognition (e.g. spiders) that is evolutionarily advantageous (Rakison & Derringer, 2008) and most researchers agree that categorical perception results from an innate mechanism. Everyday life is replete with examples of learned capabilities, such as a person's ability to play a musical instrument, to master a sport, or to become a gourmet chef. The difference in opinion over whether language fits more into the innate category or the learned category arises from the debate over whether both the mechanisms and constraints are domain-specific or domain-general. Consequently, it is necessary to focus on what type of evidence best supports the claim that language is specifically innate (i.e. constrained by a language-specific mechanism and perhaps a genetically-encoded evolutionary adaptation) or generally innate (i.e. constrained by general learning mechanisms, human perceptual and processing capacities, and production limitations).

Although nativists have acknowledged that some aspects of language can be acquired via statistical learning (e.g. lexical forms and word meanings), they maintain that other aspects of language acquisition (e.g. syntax) cannot be acquired via statistical learning. The divide between what can be learned via statistical learning and what can't falls neatly into one of two categories:

what is not a cross-linguistic universal and what is, respectively (Hauser, et al., 2002). Such criticisms stem from infants' seemingly innate preferences for linguistic input over other types of auditory input (Vouloumanos & Werker, 2004) or from the alleged "poverty of the stimulus" (Chomsky, 1980). Several studies highlight the divide between what can be learned via statistical learning and what can't, although very often the difference lies in the interpretation of the results, as opposed to the predictions. For example, arguing from the nativist perspective, Zamuner (2003) claims that the universality of CV as the unmarked syllable structure is due to an innate predisposition to the internal representation of CV patterns. A learning theory perspective, on the other hand, sees the CV pattern as a derivative of the data, one possibly ruled by constraints on the production of young speakers of a language.

It is clear that both Universal Grammar and Learning Theory require language acquisition to be constrained, yet the theories disagree on how. The constraints on language acquisition exemplified in nativist models such as Optimality Theory are specific to language – an idea in stark contrast to the constraints proposed by learning theorists which are domain-general and operant across modalities. As many of the aforementioned examples illustrate, statistical learning is a highly transferable ability, and evidence of a constraint operant in one domain should, by necessity, be able to be found in another. The present study seeks to extend the findings of an "embeddedness constraint" on statistical learning in the visual domain (Fiser & Aslin, 2005) to the auditory domain. The embeddedness constraint refers to a phenomenon in visual statistical learning where the constituent parts (embedded units) of higher-order wholes are not a relevant part of the learner's mental representation once the higher-order whole has been learned. If language learning is governed by a domain-general mechanism and the

embeddedness constraint is also domain-general, evidence for an embeddedness constraint should be present in the auditory domain as well as in the visual domain.

In the original study, the authors exposed adult participants to various nameless, arbitrary shapes over the course of five experiments. Adults were then tested to determine if and how they had formed internal representations of the stimuli. For our purpose, we will focus on the first of these experiments, in which twelve base shapes were randomly combined to form 112 unique scenes. These scenes contained only six of the twelve shapes, and were comprised of two pairs of triplets. The triplets were intermixed such that elements from each of the two triplets neighbored each other in the visual space and no clear distinction between the target triplets could be discerned. During the test phase, participants were shown triplets that they had viewed during the familiarization phase and triplets that they had never seen before. Participants were also shown pairs of shapes that were embedded within the triplets against pairs of shapes that were not present in the visual display, and asked to identify which triplets or pairs of shapes were more familiar. Fiser and Aslin discovered that once adults had formed a mental representation of a triplet, they were unable to recall any relevant information about the embedded pairs within the triplet, hence the resulting *embeddedness* constraint.

Recalling the previous discussion on the differences between Universal Grammar and Learning Theory accounts of language acquisition, two sets of predictions regarding an embeddedness constraint on language may be made. A nativist theory of language acquisition predicts that the embeddedness constraint found in the visual domain will not be found in the auditory domain because constraints on language acquisition should be domain-specific. On the other hand, a learning theory account of language acquisition predicts that the embeddedness

constraint could be present in both the visual and auditory domains because constraints on language acquisition may be domain-general.

Experiment 1: Adult Segmentation of Linguistic Stimuli

The embeddedness constraint found in the visual modality proves very useful for chunking together information. When confronted with a complex scene where multiple sub-components of that scene must be processed very quickly, a learner relies on implicit constraints to guide her to focus to the most salient features of the scene. Because of the wealth and variety of information present at any given time in any visual scene (e.g. color, shape, texture, shadow, lighting, movement, etc.), the potential problem of the curse of dimensionality is present in the visual modality as well as the linguistic modality. Constraints, such as the embeddedness constraint, may help learners circumvent such a combinatorial explosion resulting from large amounts of complex input.

An embeddedness constraint might also prove useful in language acquisition tasks, where the learner is again faced with the daunting task of processing complex input on many levels. The many components of spoken language (e.g. phonemes, stress cues, grammatical markings, transitional probabilities, etc.) can cause a combinatorial explosion just as easily as a complex visual scene. Similar to its function in the visual modality (i.e. chunking together constituent parts of higher-order components of a scene), an embeddedness constraint in language might enable learners to segment words from the speech stream more efficiently. An embeddedness constraint in language might work by chunking together syllables with high transitional probabilities into “words” that the learner could analyze for higher-order features of language (e.g. meaning or grammatical function) at a later time. If language learning is accomplished by domain-general learning mechanisms, and if the embeddedness constraint is a domain-general

constraint, we should see the same constraints in language that were previously found in the visual domain. If language is acquired via a domain-specific mechanism, or if the embeddedness constraint is a domain-specific constraint, we predict that constraints operant in the visual domain will not apply to linguistic input.

Methods

Participants

The participants were 31 undergraduate students from Carnegie Mellon University. Fifteen were familiarized with Language 1; 16 were familiarized with Language 2. Participants received either one course credit for an introductory-level psychology course or $5 for compensation.

Stimuli

Mimicking Fiser and Aslin's (2005) visual stimuli where individual shapes were combined into higher-order triplets, our languages combined 12 distinct CV syllables into 6 words: 2 trisyllabic words (L1: *tugabu* and *bapiro*; L2: *bubida* and *ropado*), 2 bisyllabic words (L1: *bida* and *pado*; L2: *bapi* and *tuga*), and 2 monosyllabic words (L1: *koo* and *tee*; L2: *noo* and *dee*). The individual syllables were synthesized in a monotone female voice with a pitch of 220 Hz using the SoftVoice vocal synthesizer. Each syllable was synthesized in its coarticulated position in the language to achieve a more natural sounding recording. The words were concatenated in a pseudo-random order, such that the transitional probabilities between syllables within a word were always 1.0. Transitional probabilities at word boundaries were much lower (0.2). This statistical structure was identical in both Language 1 and Language 2. Individual syllables had an average duration of 330 ms. There were no pauses between syllables or words. Both Language 1 and Language 2 consisted of the six words repeated six times in pseudo-

random order with no words repeating immediately after themselves. For both conditions, the language was repeated 14 times for a total learning phase of 3 minutes, 30 seconds.

In natural speech, the final syllable of the preceding word influences the initial syllable of the following word in a process called coarticulation. To replicate this property of natural speech in our synthesized language, it was therefore necessary to note which words immediately preceded and followed the target word. For example, to synthesize the word *bida,* (preceded by *pado* and followed by *koo*), we considered the string *do**bida**koo* (spaces have been eliminated to illustrate the fact that there were no distinguishing cues to word boundaries apart from statistics in our language). To synthesize each syllable in its coarticulated position, we first synthesized the string *dobida*. This string ensures that the initial syllable in *bida* is properly influenced by both the preceding syllable and the following syllable. Next, we cut the target syllable (*bi*) and pasted it into a separate audio file. Then we synthesized the next relevant string *bidakoo*. Once again we have ensured that the target syllable from *bida* (here, *da*) is accurately influenced by the preceding syllable and the following syllable. We cut the target syllable (*da*) from the string *bidakoo* and pasted it into the audio file that was created for the coarticulated syllable *bi.* This process was repeated for every syllable (and consequently every word) in the language. Individual syllables were combined to form words, and subsequently the language, using Adobe Audition 1.0 sound editing software.

Procedure

After giving informed consent, participants engaged in a learning phase and a test phase. The learning phase consisted of a 3 minute, 30 second exposure to the synthesized language. Participants listened to a .wav recording of the language on a Sony Discman CD player over headphones. Participants were instructed to attend to the language but not to analyze it during

the learning phase. Following the learning phase, participants heard five seconds of silence before beginning the test phase. For the test phase, participants were informed that they would hear a question number (e.g. "Question 1") followed by two clips of sound. The first clip of sound corresponded to "Item 1" on the worksheet; the second to "Item 2." Participants were asked to decide which item sounded more like what they had heard during the learning phase of the experiment. At test, participants completed forty 2AFC questions. There were two kinds of test questions. One kind of test trial asked participants to distinguish between words they had the opportunity to learn from the learning phase and items that were not words. On these trials, participants heard trisyllabic words (TP=1.0) and trisyllabic part-words (TP=0.33) and were asked to make a distinction between the two. Trisyllabic part-words were composed from the syllables that straddled word boundaries, such as the final syllable of a trisyllabic word and a full bisyllabic word (e.g. *bapiro* + *pado* → *ropado*). These eight of the forty questions were designed as a control to ensure that the participants had successfully segmented the words of the language.

In the second kind of test trial, participants were asked to discriminate between words and syllable sequences that had been embedded within longer words. On these trials, participants heard bisyllabic embedded words (TP = 1.0), composed from the constituent syllables of the trisyllabic words (e.g. *gabu* from *tugabu*), and bisyllabic part-words (TP=0.33) and were asked to make a distinction between the two. Bisyllabic part-words were composed from the syllables that straddled word boundaries, such as the final syllable of a trisyllabic word and the initial syllable of a bisyllabic word (e.g. *tugabu* + *pado* → *bupa*). These trials were designed to determine whether or not learners utilize an embeddedness constraint when segmenting words from the speech stream. If statistical learning of speech sequences is

governed by such a constraint, we predict that participants will perform at chance on the embedded word trials because an embeddedness constraint predicts that learners should have no knowledge of the constituent parts of a triplet they have learned.

Results and Discussion

On the trisyllabic word versus part-word control trials, participants were successful. Of the 31 participants tested, the mean score was a 64.1% (SE = 4%) correct response. A t-test indicated that this performance was significantly above chance: $t(30) = 1273$, $p < .05$. On the bisyllabic embedded word versus part-word trials, participants were also quite successful. Of the 31 participants tested, the mean score was a 60.8% (SE = 3%) correct response. This was also significantly above chance: $t(30) = 1794.2$, $p < .05$. For both kinds of test trials, chance performance was equal to a 50% correct response. Unlike Fiser and Aslin's (2005) visual study, our participants learned the embedded words of the language. When performance on the trisyllabic word trials was compared to the bisyllabic embedded word trials, there was no significant difference between the two: $F(1, 29) = 0.78$, $p = 0.39$.

These results do not indicate the presence of an embeddedness constraint. Because participants are capable of distinguishing embedded words from part-words, it is clear that they have learned something about the transitional probabilities within the embedded words. This knowledge of the transitional probabilities of components of words would not be possible had learning been guided by an embeddedness constraint. In fact, because performance on the embedded word trials was nearly as good as word trials, there may be no embeddedness constraint to overcome. The present experiment does indicate that relying only statistical cues to word boundaries, adults are capable of distinguishing between words with high transitional probabilities (TP = 1.0) and part-words (TP = 0.33). This demonstrates that adults can use

statistical learning to track transitional probabilities across words of varying syllable lengths. Although the words in our languages were trisyllabic, bisyllabic, and monosyllabic, this did not diminish adults' ability to learn both words and embedded words, suggesting that the mechanism of statistical learning is sensitive to complex linguistic input.

Experiment 2: Adult Segmentation of Tone Stimuli

To investigate the existence of the embeddedness constraint cross-modally, and to address concerns of a domain-specific linguistic adaptation for sequential auditory input versus simultaneous visual input, we conducted identical auditory experiments with tone stimuli. Should the embeddedness constraint be present in neither linguistic nor tone stimuli, it may imply that the constraint is domain-specific and adapted to simultaneous visual input. The absence of an embeddedness constraint in both linguistic and tone stimuli suggests that although the mechanism of statistical learning is domain-general, the embeddedness constraint may be domain-specific. While such an outcome may not completely differentiate between Universal Grammar and Learning Theory accounts of language acquisition, it can offer valuable insight to specific similarities and differences across modalities.

Methods

Participants

The participants were 27 undergraduate students from Carnegie Mellon University. Thirteen were familiarized with Language 1; 14 were familiarized with Language 2. Participants received either one course credit for an introductory-level psychology course or $5 for compensation.

Stimuli

The tone stimuli for Experiment 2 conformed to the same statistical design as the linguistic stimuli for Experiment 1. In this experiment, the 12 distinct CV syllables that formed the basis for the 6 words were replaced with 12 distinct tones from the chromatic scale (standard pitch; A = 440Hz). The tones ranged from C = 261.626 Hz to B = 493.883 Hz. The 12 distinct tones were combined into 6 High Probability sequences: 2 triplets (L1: *CFE* and *ADB*; L2: *ED#G* and *BF#C#*), 2 pairs (L1: *D#G* and *F#C#*; L2: *AD* and *CF*), and 2 single tones (L1: *G#* and *A#*; L2: *A#* and *G#*). The individual tones were synthesized in isolation using Adobe Audition 1.0 sound editing software. Each tone had an exact duration of 330 ms. Individual tones were combined to form sequences, and subsequently the language, also using Adobe Audition 1.0. There were no pauses between individual tones or tone sequences. Both Language 1 and Language 2 consisted of the six High Probability sequences repeated six times in pseudo-random order with no sequences repeating immediately after themselves. For both conditions, the language was repeated 14 times for a total learning phase of 3 minutes, 30 seconds.

Procedure

After giving informed consent, participants engaged in a learning phase and a test phase. The learning phase consisted of a 3 minute, 30 second exposure to the synthesized tone language. Participants listened to a .wav recording of the language on a Sony Discman CD player over headphones. Participants were instructed to attend to the language but not to analyze it during the learning phase. Following the learning phase, participants heard five seconds of silence before beginning the test phase. For the test phase, participants were informed that they would hear a question number (e.g. “Question 1”) followed by two clips of sound. The first clip of sound corresponded to “Item 1” on the worksheet; the second to “Item 2.” Participants were asked to decide which item sounded more like what they had heard during the learning phase of

the experiment. At test, participants completed forty 2AFC questions. There were two kinds of test questions. One kind of test trial asked participants to distinguish between High Probability sequences they had the opportunity to learn from the learning phase and items that were Low Probability sequences. On these trials, participants heard High Probability triplets (TP=1.0) and Low Probability triplets (TP=0.33) and were asked to make a distinction between the two. Low Probability triplets were composed from the notes that straddled sequence boundaries, such as the final tone of a tone triplet and a full tone pair (e.g. *CFE+ D#G → ED#G*). These eight of the forty questions were designed as a control to ensure that the participants had successfully segmented the High Probability sequences of the language.

In the second kind of test trial, participants were asked to discriminate between Low Probability sequences and High Probability sequences that had been embedded within longer High Probability sequences. On these trials, participants heard a sequence of two tones embedded within a High Probability triplet (TP = 1.0) and two tones that straddled sequence boundaries to form a Low Probability pair (TP=0.33) and were asked to make a distinction between the two. Low Probability pairs were composed from the tones that straddled sequence boundaries, such as the final tone of a High Probability triplet and the initial tone of a High Probability pair (e.g. *CFE + D#G→ ED#*). These trials were designed to determine whether or not learners utilize an embeddedness constraint when segmenting tone sequences from a fluent stream of tones. If statistical learning of tone sequences is governed by such a constraint, we predict that participants will perform at chance on the embedded word trials because an embeddedness constraint predicts that learners should have no knowledge of the constituent parts of a triplet they have learned.

Results and Discussion

On the High Probability triplets versus Low Probability triplets control trials, participants were successful. Of the 27 participants tested, the mean score was a 57.4% (*SE* = 5%) correct response. A t-test indicated that this performance was significantly above chance: $t(26) = 1101.2$, $p < .05$. On the High Probability embedded pairs versus Low Probability pairs trials, participants were also successful. Of the 27 participants tested, the mean score was a 57.6% (*SE* = 2%) correct response. This was also significantly above chance: $t(26) = 2311.4$, $p < .05$. For both kinds of test trials, chance performance was equal to a 50% correct response. Similar to Experiment 1, and once again unlike Fiser and Aslin's (2005) visual study, our participants learned the High Probability embedded pairs of the language. When performance on the triplet trials was compared to the pair trials, there was no significant difference between the two: $F(1, 25) = 0.01$, $p = 0.93$.

A comparison of adults' performance in Experiment 1 to adults' performance in Experiment 2 reveals no significant difference between speech and tone stimuli. Of the 58 participants tested in both experiments, the mean score was a 61% correct response on words/High Probability triplets (TP=1.0). On the embedded words/High Probability pairs (TP=1.0), the mean score was a 59.3% correct response for the 58 participants tested. When performance on test items (triplets versus pairs) was compared across experiments, there was no significant difference between the two: $F(1, 56) = 0.26$, $p = 0.61$. Moreover, when performance on test items (triplets versus pairs) with respect to stimulus (speech versus tone) was compared across experiments, there was no significant difference between the two: $F(1, 56) = 0.35$, $p = 0.56$.

These results do not indicate the presence of an embeddedness constraint. Because participants are capable of distinguishing High Probability embedded pairs words from Low

Probability pairs, it is clear that they have learned something about the transitional probabilities within the embedded sequences. This knowledge of the transitional probabilities of components of sequences would not be possible had learning been guided by an embeddedness constraint. In fact, because performance on the embedded word trials was nearly as good as word trials, there may be no embeddedness constraint to overcome. The present experiment does indicate that relying only statistical cues to sequence boundaries, adults are capable of distinguishing between High Probability triplets (TP = 1.0) and Low Probability (TP = 0.33). Adults successfully segmented the High Probability triplets despite the varying length of the sequences, indicating a statistical learning mechanism that is highly sensitive to variable input.

The finding that the embeddedness constraint is absent in both linguistic and tone stimuli may suggest that the constraint is adapted to simultaneous visual input rather than sequential auditory input. The absence of an embeddedness constraint in both types of auditory stimuli suggests that although the mechanism of statistical learning is domain-general, the embeddedness constraint may be domain-specific. While this result does not facilitate differentiation between Universal Grammar and Learning Theory accounts of language acquisition, it does offer valuable insight to a potentially unique and subtle difference across modalities.

Experiment 3: Infant Segmentation of Tone Stimuli

Finally, we have begun a cross-age comparison of statistical learning of both speech and tone sequences in order to more rigorously test hypotheses of domain-generality or domain-specificity as well as to investigate a possible change in the mechanism over time. Most Learning Theory accounts of language acquisition, particularly those that emphasize the domain-general ability of statistical learning, do not predict a difference in performance between infant

and adult populations on word segmentation tasks. However, nativist accounts of language acquisition predict a difference between infant and adult populations because of the domain-specificity of the mechanism (i.e. a special Language Acquisition Device) as well as a maturational change (i.e. the disappearance of the LAD over time). The results of our infant testing, in conjunction with our experiments on adults, should provide persuasive evidence for one of the two accounts of language acquisition.

Methods

Participants

The participants were 20 infants ranging in age from 13.5 months to 14.5 months. Participants had a mean age of 14.15 months. There were 7 participants in Language 1 (3 male, 4 female) and 15 participants in Language 2 (8 male, 7 female). In order to obtain the 20 participants for this experiment, it was necessary to test 22. Two participants were excluded from Language 2 for the following reasons: fussy (2). Participants received either a book or a t-shirt for compensation. All infants had normal hearing and were free from ear infections at the time of testing, according to parental report.

Stimuli

The tone stimuli for Experiment 3 were identical to the tone stimuli from Experiment 2.

Procedure

After receiving a parent or legal guardian's informed consent, participants engaged in a learning phase and a test phase. Participants were tested individually in a sound-attenuated room while seated on a parent's lap, 150 cm away from a 30" Apple cinema display monitor. Two speakers presented the audio stimuli, one on either side of the display. An experimenter presented the stimuli and coded looking times online using Habit X (Cohen, Atkison, and

Chaput, 2004). The experimenter observed the child's looking time over a closed-circuit monitor from a separate room while blind to the stimuli. During the leaning phase, coders initiated presentation of the stimuli but did not measure looking time. Following the learning phase, an attention-getting animated film of Winnie the Pooh, coupled with a recorded verbal encouragement, was used to fixate the infant's view on the monitor. Coders used the infant's initial attention to the monitor during the attention-getting film as a baseline to determine whether infants were looking at the monitor or looking away from it. Using this method, when coders determined that the infant was looking at the monitor, they pressed a key for the duration of the child's attention to the monitor. When the infant looked away from the monitor, coders released the key until the child re-oriented to the screen or until the attention-getting film began to play.

The experiment consisted of a learning phase and a test phase. The learning phase was 4 minutes long. During the learning phase, the tone language was presented over the speakers and participants saw a black-and-white checkerboard on the display. Following the learning phase, the test phase began with an attention-getting animated film of Winnie the Pooh, coupled with a recorded verbal encouragement. The test phase consisted of 6 randomized trials: 3 High Probability tone triplets (TP=1.0) and 3 Low Probability tone triplets (TP=0.33). Once the child's attention was fixed on the display, the experimenter presented the first, randomized test item. An image of a looming, green ball appeared on screen as the test stimulus began to play. The stimulus repeated for 25 seconds or until the participant looked away for more than 2 seconds. The tonal triplets presented at test were composed exactly like the tone triple foils from Experiment 2 with adults. However, infants were only presented with two triplets (TP=1.0 and TP=0.33). Infants exposed to Language 1 were presented with the High Probability triplet *CFE*

and the Low Probability triplet *BF#C#*. Infants exposed to Language 2 were presented with the High Probability triplet *ED#G* and the Low Probability triplet *ADB*. In the test phase for both languages, the High Probability triplet and the Low Probability triplet were each repeated 3 times, for a total of six test trials. In between each of the 6 test trials, participants viewed the attention-getting film.

Results and Discussion

Due to the currently incomplete sample (our goal is to run 32 more infants), none of the results have reached significance. However, the results are potentially consistent with successful segmentation. On the word segmentation task (High Probability versus Low Probability trials), infants listened to High Probability test items for 13.5 seconds and to Low Probability items for 14.5 seconds. The magnitude of this difference – 1 second – is consistent with previous experiments that have demonstrated successful learning, with larger sample sizes (e.g., Saffran et al., 1996). Across languages, there was no significant difference between looking times for High Probability and Low Probability items: $F(1, 18) = 1.6$, $p = 0.22$. A comparison of test items with respect to stimulus also did not indicate any significant difference between Language 1 and Language 2: $F(1, 18) = 0.3$, $p = 0.61$.

While these results are preliminary, they suggest that infants may be able to learn the languages presented. However, the inequality in number of participants to condition (7 participants were exposed to Language 1, 13 to Language 2) must be corrected, and 6 more participants run to balance the two conditions. If a larger sample size yields significant discrimination between High and Low probability sequences, this experiment will be the first to demonstrate that infants can identify word boundaries in a statistical language consisting of varied sequence lengths (tone triplets, pairs, and single tones). Pending positive results with the

segmentation task, we plan to test infants on the embeddedness constraint in a manner consistent with adult testing.

General Discussion

Although the present study does not provide evidence for an embeddedness constraint operant on auditory input, we have demonstrated that adults (and potentially infants) can successfully segment a language that contains words of varying lengths (i.e. monosyllabic, bisyllabic, and trisyllabic words) using only statistical cues to word boundaries. We have also shown that adults can successfully segment a tone language that contains tone sequences of varying lengths (triplets, pairs, and single tones). In using auditory input with high probability sequences of varying lengths we have taken a significant step toward the goal of increasing the complexity of artificially-generated statistical languages. With respect to the speech stimuli, by making our statistical languages more complex and more similar to natural language, we hope to counter arguments that statistical learning of word boundaries is irrelevant outside a laboratory setting (see Johnson & Seidl, 2008; Yang, 2004). Our results with infants did not provide conclusive evidence that infants were capable of using statistical cues to segment a complex tone language. However, further testing with infant participants using the same tone stimuli (as well as ongoing experiments with speech stimuli) will indicate whether or not there is a difference in infants' ability to learn one type of auditory input over another.

There are several compelling hypotheses we can derive from the current lack of evidence for an embeddedness constraint on auditory input. First, it is unclear whether the embeddedness constraint is unique to the visual domain or whether the embeddedness constraint applies to simultaneous, rather than sequential, input. In future experiments, it may be informative to present adults with a sequential display of the same visual stimuli found in Fiser and Aslin's

(2005) original study. If the embeddedness constraint is not present during the sequential display of the same shapes from the original study, we can expect that the embeddedness constraint applies to the manner, as opposed to the modality, in which the input is presented. Second, it is possible that the embeddedness constraint decays with extensive experience with the input. The absence of the constraint in adult learners could be the product of vast experience with language. English is not a morphologically rich language, but it nevertheless contains many examples of meaningful syllable sequences embedded within words (e.g. taste versus taste*ful* versus *dis*taste*ful*). The prior knowledge that constituent parts of words are important could override an embeddedness constraint that demands otherwise. Pending studies with infants and linguistic stimuli will confirm or deny this hypothesis. However, as previously mentioned, our experiments with adults and tone stimuli do not indicate that an embeddedness constraint governs segmentation of tone sequences. As our adult participants had no extensive musical experience or training, we believe this finding suggests that there may be no embeddedness constraint to override. Therefore, experience with the input may not be as important as the manner in which the input is presented (i.e. sequential versus simultaneous). The results of more rigorous testing of experience versus manner, along with testing of modality versus manner, should provide evidence that will inform us about the nature of the embeddedness constraint. Taken together with results indicative of an embeddedness constraint that is present or absent during infancy, these experiments could provide a new perspective on the evolution of the statistical learning mechanism over time.

Other potential reasons for the lack of evidence for the embeddedness constraint in auditory input stem from the input itself. Although both our speech and tone languages were more complex than other statistical languages explored in the literature, they were far less

complex than the visual stimuli from the original Fiser and Aslin (2005) study. Perhaps most notably, the components of the triplets in Fiser and Aslin's study were promiscuous and occurred in multiple settings. In our experiment, each component only ever occurred in one word or sequence. Future experiments with less fixed component-to-sequence relationships may provide evidence for an embeddedness constraint in the auditory domain. Possibly another significant difference between Fiser and Aslin's (2005) visual stimuli and our auditory stimuli was the number of times participants were exposed to any given triplet. In the original experiment with visual stimuli, participants viewed each of the four High Probability triplets fifty-six times. This gave participants fifty-six exposures to an item that might have something embedded within it. In our experiments, participants heard each of the two High Probability triplets six times per iteration of the language. From this perspective it is possible that adults have been over-exposed to the visual triplets and under-exposed to the auditory triplets. Preliminary testing of our linguistic input with the PARSER model of word segmentation (Perruchet & Vinter, 1998) suggests that knowledge of embedded segments decreases or vanishes with increased exposure. More rigorous testing with both PARSER and human participants is planned to determine whether or not this hypothesis will prove valid.

Although the powerful ability of statistical learning has been the object of extensive research in recent years, the mechanisms underlying statistical learning still are not fully understood (Perruchet & Pacton, 2006; Conway & Christiansen, 2005). In particular, three different theories exist to explain how statistical learning occurs: associative theories, Bayesian theories, and chunking theories. Associative theories of statistical learning claim that units from the input are learned one at a time by computing statistical relations between the units. The individual units are then combined, bottom-up, to form a cohesive whole. Bayesian theories of

statistical learning suggest that units from the input are considered with respect to their prior and posterior probabilities; for example, a prospective unit (A) is assessed against the likelihood of other prospective units (B, C, etc.). The most consistent units become the most preferred with less consistent units falling out of favor (see Chater et al., 2006 for discussion). Chunking theories of statistical learning hypothesize that chunks are generated from working memory and processing limitations. Chunks are then strengthened or weakened via associative memory. Exploring the embeddedness constraint across domains has the potential to provide strong domain-general evidence for either an associative, Bayesian, or chunking theory of statistical learning.

The embeddedness constraint is well suited to fit within a Bayesian theory of language acquisition because Bayesian learning rests on the assumption that the most narrow and consistent hypothesis is preferred. However, if narrowness and consistency are in direct competition, consistency trumps narrowness. In a comparison related to our experiments, embedded words are always associated with a third syllable (that of the trisyllabic word in which they are embedded), making them less consistent than bisyllabic words, and, according to a Bayesian account of statistical learning, causing embedded words to cease to be considered. The embeddedness constraint is effectively irrelevant to an associative learning theory, which relies on a bottom-up composition of words. When confronted with two wholes of equally predictive statistics, learners should feel comfortable selecting either one, and thus embedded words and bisyllabic words are equally acceptable to the associative model. Interestingly, chunking may be consistent with evidence for an embeddedness constraint as well as the lack thereof. As previously mentioned, preliminary testing of linguistic input with PARSER (Perruchet & Vinter, 1998) indicates that with much exposure knowledge of embedded segments vanishes. However,

with less exposure, embedded segments are considered to be more viable candidates for words than the triplets in which they are embedded. This may suggest that associative memory, which functions to strengthen or weaken chunks, operates differently at different points in time during the learning phase.

Research investigating the existence of the embeddedness constraint may not only produce the data to disambiguate between three competing theories for a mechanism of statistical learning, it may also indicate that different mechanisms guide learning in different domains or at different points in the lifespan. One of the foremost goals of modern psychology is to define the powerful mechanisms underlying human cognition (Thiessen, 2009; Marcus, et al., 1999; Bloom & Markson, 1998). It is with this in mind that we have begun to pursue a line of cross-modal, cross-age experiments to investigate the existence of the embeddedness constraint on linguistic input.

Works Cited

Bickerton, D. (1981). *Roots of Language.* Karoma Publishers.

Bloom, P. & Markson, L. (1998). Capacities Underlying Word Learning. *Trends in Cognitive Sciences, 2,* 67-73.

Chater, N., Tenenbaum, J., & Yuille, A. (2006). Probabilistic Models of Cognition. *Trends in Cognitive Sciences, 10,* 287-291.

Chomsky, N. (1959). A Review of B. F. Skinner's *Verbal Behavior*. In Leon Jakobovits and Murray S. Miron (eds.), *Readings in the Psychology of Language,* Prentice Hall, 142-143.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* London: The MIT Press.

Chomsky, N. (1980). *Rules and Representations.* Oxford: Basil Blackwell.

Christiansen, M. H. & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences, 31 (5),* 489-509.

Cohen, L. B., Atkison, D. J., & Chaput, H. H. (2004). Habit X: a new program for obtaining and organizing data in infant perception and cognition studies (Version 1.0). Austin: University of Texas.

Comrie, B. (1989). *Language Universals and Linguistic Typology*. University of Chicago Press.

Conway, C. M. & Christiansen, M. H. (2005). Modality-Constrained Statistical Learning of Tactile, Visual, and Auditory Sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 24-39.

Conway, C. M. & Christiansen, M. H. (2006). Statistical Learning Within and Between Modalities: Pitting Abstract against Stimulus-Specific Representations. *Psychological Science, 17 (10),* 905-912.

Culicover, P. W. (1997). *Principles and Parameters: An Introduction to Syntactic Theory.*

Oxford: Oxford University Press.

Fiser, J. & Aslin, R. N. (2001). Unsupervised Statistical Learning of Higher-Order Spatial Structures from Visual Scenes. *Psychological Science, 12,* 499-504.

Fiser, J. & Aslin, R. N. (2005). Encoding Multielement Scenes: Statistical Learning of Visual Feature Hierarchies. *Journal of Experimental Psychology: General, 134,* 521-537.

Gold, E. M. (1967). Language Identification in the Limit. *Information and Control, 10,* 447-474.

Hauser, M.D., Newport, E.L., & Aslin, R.N. (2001). Segmentation of the speech stream in a nonhuman primate: Statistical learning in cotton top tamarins. *Cognition, 78,* B53-B64.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science's Compass, 298,* 1569-1579.

Johnson, E. K. & Seidl, A. H. (2008). At 11 Months, Prosody Still Outranks Statistics. *Developmental Science, 11,* 1-11.

Jusczyk, P., & Aslin, R. (1995). Infants' Detection of the Sound Pattern of Words in Fluent Speech. *Cognitive Psychology*, *29*, 1-23.

Kager, R. (1999). *Optimality Theory.* Cambridge University Press, Cambridge.

Kielar, A., Joanisse, M. F., & Hare, M. L. (2008). Priming English Past-Tense Verbs: Rules or Statistics? *Journal of Memory and Language, 58,* 327-346.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual Statistical Learning in Infancy: Evidence for a Domain General Learning Mechanism. *Cognition, 83,* B35-B42.

Landau, B. (1998). Nativist perspectives on the acquisition of knowledge. In W. Bechtel & G. Graham (Eds.), *A companion to cognitive science.* Oxford, UK: Blackwell.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule Learning by Seven-

Month-Old Infants. *Science, 283,* 77-80.

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

Newport, E. L. (1990). Maturational Constraints on Language Learning. *Cognitive Science, 14,* 11-28.

Perruchet, P. & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language, 39,* 246-263.

Perruchet, P. & Pacton, S. (2006). Implicit Learning and Statistical Learning: One Phenomenon, Two Approaches. *Trends in Cognitive Sciences, 10,* 233-238.

Pinker, S. (1994). *The Language Instinct: How the Mind Creates Language.* Harper Collins, 2000.

Prince, A. & Smolensky, P. (2004) *Optimality Theory: Constraint Interaction in Generative Grammar.* Blackwell Publishing.

Quine, W. V. O. (1960). *Word and Object.* The MIT Press, Cambridge MA.

Rakison, D. H. & Derringer, J. L. (2008). Do infants possess an evolved spider-detection mechanism? *Cognition, 107,* 381-393.

Saffran, J., Aslin, R., & Newport, E. (1996a). Statistical Learning by 8-month-old Infants. *Science, 274,* 1926-1928.

Saffran, J., Newport, E., & Aslin, R. (1996b). Word Segmentation: The Role of Distributional Cues. *Journal of Memory and Language, 35,* 606-621.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical Learning of Tone Sequences by Adults and Infants. *Cognition, 70,* 27-52.

Smith, N. (2005). Chomsky's science of language. In J. McGilvray (Ed.) *The Cambridge*

Companion to Chomsky. Cambridge: Cambridge University Press.

Thiessen, E. D. (2009). Statistical Learning. E. Bavin (ed.), *Cambridge Handbook of Child Language.*

Vouloumanos, A. & Werker, J. F. (2004). Tuned to the signal: the privileged status of speech for young infants. *Developmental Science, 7,* 270-276.

Yang, C. D. (2004). Universal Grammar, Statistics, or Both? *Trends in Cognitive Sciences, 8,* 451-456.

Zamuner, T. S. (2003). *Input-based Phonological Acquisition.* New York: Routledge.