TEMPORAL EXPECTANCY AND THE EXPERIENCE OF STATISTICS IN LANGUAGE PROCESSING

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAI‘I AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

LINGUISTICS

August 2010

By

Hunter B. Hatfield

Dissertation Committee:

William O’Grady, Chairperson
Victoria Anderson
Kamil Ud Deen
Luca Onnis
Amy Schafer
Susanne Still

Keywords: Time, Statistical Learning, Prosody, Word Segmentation
ACKNOWLEDGEMENTS

A curious feature of dissertations is that the almost exclusive readers for such works are colleagues and fellow researchers in the academic community. However, typically the person most responsible for the accomplishment of the dissertation is not connected through an academic relationship, but a personal one. In this case, I must thank my wife, Nathalie George. When I approached her saying something like, “I’d like to quit my corporate job, sell our house and virtually all our possessions, move to a state I’ve never even visited with our 1-year-old, and live in a 1 bedroom apartment again as if we were still 23, all so that I might have a slim possibility of a job in research,” she might have justifiably said, “maybe not.” But she said, “okay,” and this dissertation before you is the result.

I must also thank the members of my committee, starting with my advisor, William O’Grady. He managed to stick with the topic through thick and thin, despite the false starts and dead ends. What started off as a discussion about statistical learning in Korean turned into an essay on metrical phonology for English-speaking adults. My other committee members have all contributed in various ways, most importantly through my education over a 5-year period. If you take what they have taught me, mix it around, stick in my tendency towards hopefully useful speculation, and stir in a little noise from my fuzzy brain, you end up with something like the ideas within here. Thank you to Victoria Anderson, Kamil Deen, Luca Onnis, Amy Schafer, and Susanne Still.

Of course, classmates and colleagues helped as sounding boards, put-upon pilot experiment participants, sanity checks, 1AM lab company, and more. Much thanks to
Junghee Kim, Julia Wieting, Bodo Winter, the ideal colleague Jinsun Choe, Kaori Ueki, Manami Sato, and many now-insulted friends.

And, finally, thanks to the one-year-old who is now finishing first grade. He had to put up without a daddy for much of the month of April and deserves credit.
ABSTRACT

This dissertation assesses the relationship between statistical learning and temporal perception. It starts entertaining a bold hypothesis: that former demonstrations of statistical learning were actually demonstrations that isochronous word onsets could be used to segment words within speech. To assess this, two languages are created. One language employs varying word lengths (2 and 3 syllables) and varying word durations. The second employs varying word lengths and identical word durations. It is expected that learning will be better in the case with identical word durations.

Three conclusions are reached through analysis of the resulting data. 1) The data cannot be adequately explained without positing knowledge of the statistical distributions of syllables. This then rejects the hypothesis that isochronous word onset intervals created a confound in previous work. However, the statistical knowledge is most consistent with the notion that the distributional patterns are signaling a prosodic break, not a lexical one. The Information / Duration hypothesis is presented along with this argument. This hypothesis states that an increase in uncertainty will be experienced as an increase in duration. 2) The time course of word segmentation should not be overlooked. Previous claims that one cue is stronger for segmentation than another cue might be better explained by temporal priority. Cues that are encountered first will set expectations more than later cues. 3) Statistical learning should result in greater demonstrations of learning than seen in the experimental results. This is most consistent with the presence of a competing cue. Entrainment to a rhythmic stimulus, the earlier proposed confound, is the most natural competing cue here. Much of the work is interpreted within theories of time perception based upon dynamic oscillators.

The main result is that attention is a prime mechanism to control what sorts of items are calculated in statistical learning, and rhythm is one method to control attention. The dissertation also assesses what it is like for a speaker to experience a statistical distribution rather than simply calculate it.
# TABLE OF CONTENTS

Acknowledgements .......................................................... ii
Abstract ................................................................. v
Table of Contents ......................................................... v
List of Tables ............................................................. vi
List of Figures ............................................................ vii
Chapter 1 ....................................................................... 1
Chapter 2 ....................................................................... 25
Chapter 3 ....................................................................... 44
Chapter 4 ....................................................................... 61
Chapter 5 ....................................................................... 73
Chapter 6 ....................................................................... 101
Chapter 7 ....................................................................... 105
Chapter 8 ....................................................................... 135
Bibliography .................................................................... 148
LIST OF TABLES

Table 1. Words Distinguished by Stress in English (stress indicated with CAPS) ........................................ 29
Table 2. Summary of Word Segmentation Experiments. ................................................................. 56
Table 3. L-Words and Boundary Words for Experiment 1 ............................................................. 64
Table 4. L-Words and Boundary Words for Experiment 2 ............................................................. 73
Table 5. Experiment 2A Selected Words .......................................................................................... 80
Table 6. Mixed Effects Model of Voiceless Stops as Predictors of Accuracy .................................. 81
Table 7. Experiment 2A Selected Words Grouped by Word Pairings ........................................... 83
Table 8. Counts of Word Selection Without Training ........................................................................ 84
Table 9. Counts of Word Selection Without Training Grouped by Language .................................. 85
Table 10. Experiment 2B Selected Words ....................................................................................... 96
Table 11. Mixed Effects Model of Voiceless Stops as Predictors of Accuracy ................................. 96
Table 12. Experiment 2B Selected Words Grouped by Word Pairings .......................................... 97
Table 13. Words and Boundary Words for Experiments 2A and 2B .............................................. 98
Table 14. Experiment 2D Selected Words ....................................................................................... 103
Table 15. Experiment 2D Selected Words Grouped by Word Pairings .......................................... 104
Table 16. Experiment 3A Selected Words ....................................................................................... 116
Table 17. Experiment 3A Selected Words Grouped by Word Pairings .......................................... 117
Table 18. Disyllabic l-words Followed by Trisyllabic L-Words ....................................................... 118
Table 19. Experiment 3B-Low Selected Words ............................................................................ 126
Table 20. Experiment 3B-Low Selected Words Grouped by Word Pairings .................................... 127
Table 21. Experiment 3B-High Selected Words ............................................................................... 130
Table 22. Experiment 3B-High Selected Words Grouped by Word Pairings .................................. 130
LIST OF FIGURES

Figure 1. Tilsen’s (Submitted) model of oscillator-based gestural control. (Reproduced from Tilsen (Submitted, Figure 1) by permission.) . . . . . . . . . . 41
Figure 2. A syllable-level oscillation repeats three times for each foot-level oscillation. . 47
Figure 3. Mean proportion correct with confidence intervals for experiment 1. . . . . . 66
Figure 4. Density curves for experiments 1A and 1B. . . . . . . . . . . . . . . 67
Figure 5. Histogram of proportion correct for Experiment 1. . . . . . . . . . . . 70
Figure 6. Mean proportion correct with confidence intervals for experiment 2. . . . . 75
Figure 7. Density curves for experiments 2A and 2B. . . . . . . . . . . . . . . 75
Figure 8. Mean proportion correct with confidence intervals for experiments 1 and 2. . 76
Figure 9. Density curve for experiment 2D. . . . . . . . . . . . . . . . . . . . 102
Figure 10. Density curves for experiments 3A and 3B. . . . . . . . . . . . . . . 107
Figure 11. Histogram plots for experiments 3A and 3B. . . . . . . . . . . . . . . 107
Figure 12. Density curves for experiments 3B-High and 3B-Low. . . . . . . . . . 108
Figure 13. Mean proportion correct for experiments 1, 2, and 3. . . . . . . . . . 111
Figure 14. Two coupled oscillators coupled to produce three metrical feet. . . . . . 133
Figure 15. A 300 ms. oscillator with syllabic input . . . . . . . . . . . . . . . . . 145
Figure 16. A 600 ms oscillator coupled with a 300 ms oscillator . . . . . . . . . 146
CHAPTER 1. SETTING THE PROBLEM

1.1. Statistical Learning and Statistical Processing

One of the hottest topics in all of linguistics, and certainly language acquisition, over the last 15 years has been that of statistical learning; i.e., learning language from the statistical patterns inherent in what we hear and see. In 1996, Jenny Saffran, Richard Aslin, and Elissa Newport published two papers, one concerning adults (Saffran, Newport, & Aslin, 1996) and another concerning infants (Saffran, Aslin, & Newport, 1996), showing that people could use the statistical structure of what they hear to solve a linguistic problem such as finding words in speech. Unlike many written forms of language, such as the text you are reading now, natural speech does not typically place a space or pause between each word. Words in the acoustic stream of spoken speech do not have clear physical breaks separating them but instead flow together with full pauses only coming at the end of larger phrases or entire utterances.

This becomes entirely obvious whenever one encounters an unknown language. Not only do we not know what the words mean in these unfamiliar languages, we do not know what the words are. However, once a language is known, we appear to hear a series of words. The average person can be tested in school on how to spell words; linguists can contemplate their structure and analyze how they are put together into sentences. Somehow, an infant must go from this state of a buzzing, blooming confusion of streaming speech to segmented language. How do we do it?

The two papers from Saffran, Aslin, and Newport provided one possible answer. We can count the syllables as they occur. Syllables that cluster together are more likely to be inside words in the same way as ingredients in a common dish are more often found
together: We are more likely to see tomato sauce if we have seen spaghetti than we are to see mangoes if we have seen spaghetti. This intuition that pieces of language occur together can be formalized with the notion of transitional (or conditional) probability.

Transitional probability is the chance of one item occurring given the presence of some other item. Given spaghetti, one is more likely to see a tomato product than a mango. Given the syllable ba- [bei], we are more likely to next hear the syllable –by [bi] than the syllable –sprat [spræt], because baby is a word of English, while basprat is not. Such a metric will not always work to find the boundaries of words, but probabilistic cues do work surprisingly well. Corpus analysis has revealed that the distribution of sounds in American English is highly bimodal with the majority of phoneme pairs occurring exclusively either within words or across word boundaries (Hockema, 2006). Several different models have demonstrated relative success finding words in English with statistical computation (Brent, 1999; Brent & Cartwright, 1996; Swingley, 2005), though segmentation naturally improves if additional cues are used (e.g., Yang, 2004). Indeed, for at least 50 years before the Saffran papers, the idea that statistical computation could be useful to understand linguistic knowledge had been proposed and debated (Harris, 1955; Shannon, 1948). There are useful probabilistic cues to word boundaries in natural language if humans have the ability to use them. This is where the Saffran et al. papers come in. They provide evidence that both 8-month-old infants and adults can indeed use such cues.

1.2. Methodology of Saffran, Newport, and Aslin (1996)

How exactly did Saffran et al. provide evidence of statistical learning? The key idea was to create an artificial “language” such that the words of the made-up language
could only be found through a statistical calculation. Saffran, Newport, and Aslin (1996), which worked with adults, created an artificial language composed of six trisyllabic words: *babupu, bupada, dutaba, patubi, pidabu*, and *tutibu*. These words were then semi-randomly concatenated with no pauses or other known cues between words into three blocks of seven minutes each. The only restriction to the randomization was that one word could not repeat itself. Orthographically, the speech stream would be something like (1).

(1) bupadatutibabupupatubibupadabupupidabututibu…

A synthesizer produced the speech with a flat intonational pattern, constant intensity, and identical syllable durations, which is intended to remove prosodic cues from the speech.

After listening to the artificial speech in a training session, adults were presented with several pairs of words. One test included a word of the language and a non-word (a word composed of syllables that never occurred together during training). Another test included a word and a part-word (a word composed of two syllables that occurred within a word plus a third syllable from a different word).

(2) Example of Word vs Non-Word: bupada vs budati
(3) Example of Word vs Part-Word: bupada vs pidata

In example (2), the syllables of the non-word occur within the speech stream but never together, while in (3) the syllables *pi* and *da* are part of the word *pidabu*, while *ta* is from a different word. If the experimental design has no confounds, the only way to make such discriminations is to notice that certain syllables occurred together more frequently than others. Specifically, the transitional probability among syllables inside a word will be systematically higher than among syllables that cross word boundaries. In the words versus non-words condition, participants chose successfully 76% of the time, while, in
the more difficult part-word condition, they chose correctly 65% of the time. Both results are significantly different from the chance level of 50%.

1.3. Why is Saffran et al. (1996) Important?

1.3.1. Statistics in word segmentation

At the time of writing, Saffran, Aslin, & Newport (1996), the paper related to statistical learning in infants, has been cited approximately 1500 times, according to Google Scholar, and Saffran, Newport, & Aslin (1996), the paper related to statistical learning in adults, has been cited another 420 times. Together then, we have around 1900 citations over a 14 year span, which implies that approximately 136 papers a year refer to this work. While this work does not “top the list” of most citations (Noam Chomsky’s *The Minimalist Program*, for instance, has 9,000 citations over 15 years), in the field of linguistics, 1900 citations is quite a response.

Why has there been so much interest in these papers? Certainly, part of the interest is from researchers interested in word segmentation. There is a large body of work that examines word segmentation with both adults and infants, and much of it incorporates the results of Saffran et al. Many cues have been discovered for finding word boundaries, including stress (Johnson & Jusczyk, 2001; Jusczyk, Houston, & Newsome, 1999), phonotactic patterns (Mattys & Jusczyk, 2001), allophones (Jusczyk, Hohne, & Bauman, 1999), word final vowel lengthening, (Saffran, Newport, & Aslin, 1996), prosodic phrase markings (Kim, 2004; Shukla, Nespor, & Mehler, 2007), and co-articulation and articulatory strengthening (Johnson & Jusczyk, 2001). Besides simply including transitional probabilities as one of the cues for segmenting words, a good number of studies have directly pitted a statistical pattern against another possible cue.
Shukla et al. (2007) found that words in an artificial language that are composed in a statistical pattern are not heard if they cross an intonational phrase boundary. Similarly more than 80% of content words in English follow a trochaic strong/weak pattern in which the first syllable is stressed and the second unstressed (Cutler & Norris, 1980). If such a trochaic pattern is placed across a word boundary in an artificial language so that the stress pattern marks part-words, while the statistical pattern marks the language’s words, infants hear the part-words, not the words (Johnson & Jusczyk, 2001; Johnson & Seidl, 2009). While no one has yet directly pitted phonotactic cues to word boundaries against statistical patterns to word boundaries, we do have evidence that adults will segment words based upon phonotactic evidence over conflicting stress evidence (McQueen, 1998). If phonotactics are stronger than stress and stress is stronger than statistics, we can infer through transitivity that phonotactics will also be stronger than statistics.¹

This research is typically framed in the terms just used – one cue is stronger or has more weight than another. Johnson and Seidl (2009), for instance, advocate a program of investigating the relative rankings of segmentation cues throughout development. Greater weight is not the only reason a preference might be displayed, however. An oft-neglected aspect of statistical learning experiments is their time course. A statistical pattern in speech can only be discovered over time. One must start “counting” the syllables and discover their distribution as they are encountered in the training part of the experiment. Cues that depend upon prior knowledge, however, can be used immediately. For instance, as English speakers are aware that most words start with

¹ However, there may be changes in cue “rankings” over development. See Thiessen & Saffran (2007) and Johnson and Seidl (2009) for reviews and discussion.
a stressed syllable, they can guess that a stressed syllable is the start of a new word from the very first word they hear during an artificial language experiment. Similarly, phonotactic patterns that are present in the language spoken by the participants (for example, [h] is only found at the beginning of a syllable in English) can be used to segment words immediately. This implies that any segmentation based upon statistics must compete with existing segmentations done with cues that were available earlier than the statistical pattern was available. The result is a weight based upon temporal priority, not strength.

The methodology used for much experimentation on word segmentation does not lend itself to looking at the time course of segmentation, however. For instance, Saffran, Newport, and Aslin (1996) trained participants for 21 minutes and then tested them immediately afterwards. It is difficult to compare perception of the speech at 1 minute versus at 15 minutes under this paradigm. We might be able to see evidence for changing perceptions if evidence lingers all the way to the test. In other words, if at the beginning of an experiment a participant uses cue #1 to find a set of possible words A and later in an experiment uses cue #2 to find a set of possible words B, there might be evidence remaining in the test results that both sets were heard, but only if cue #2 does not completely remove the perception of cue #1. If cue #2 does entirely overwhelm cue #1, we would lose evidence that cue #1 was ever considered in the training-and-then-test methodology.

Previous research does provide some evidence that the time course of statistical learning is important. Gebhart, Aslin, & Newport (2009) examined how participants would segment words when the statistical structure of the language was changed partway
through training. In their work, participants first heard one language with statistical structure #1 followed by a second language with statistical structure #2. Unless an explicit cue (in the form of a 30 second pause between languages) was provided, participants only learned the first language. With the long pause between languages, they were able to learn both languages. However, in the absence of the explicit pause, if the training on the second language were triple the duration of the first, then they would learn the second. Together, this suggests that earlier segmentations of an artificial language are given precedence over later segmentation unless the latter segmentation cue is over a substantially longer period.

Such results corroborate general tendencies to process language immediately and incrementally before large samples of data are present. Of course, immediate parsing of speech has drawbacks, particularly when an error is made. Gebhart et al. express it well: “However, if learners base their structural hypotheses on small samples, it opens the door to garden-path errors, and most learning models have difficulty recovering from such errors without huge amounts of countervailing data to overcome the initial incorrect structural hypothesis” (2009, p. 1088). Applied to the case of stress and statistics, stress would be available immediately for parsing and participants are likely to use it. Even if they continue to calculate the statistical structure of the language, it could take “overwhelming” evidence for the participant to back off from the first parse and employ the second for segmentation.

1.3.2. Statistical learning beyond word segmentation

While word segmentation is certainly a significant issue in psycholinguistics, there are not 136 papers a year on this topic. Arguably, what is most important about the
Saffran et al. papers is not the implications for word segmentation specifically, but the simple fact that people can use statistics at all. If we can learn from a relatively sophisticated measure such as transitional probability, then, in theory, almost any linguistic property that has a surface statistical manifestation is now learnable.

Indeed, an explosion of research has occurred in the past 15 years that appears to show that distributional information can allow listeners to learn all sorts of linguistic properties. Statistical learning has been documented with phrase structure (Thompson & Newport, 2007), hierarchical phrase structure (Takahashi & Lidz, 2008), syntactic structure for second language acquisition (Onnis, Waterfall, & Edelman, 2008), verb argument structure (Wonnacott, Newport, & Tanenhaus, 2007), phoneme categorization (Maye, Werker, & Gerken, 2002), stress assignment (Gerken, 2004), and simple grammars (Gomez, 2002, 2006; Onnis, Monaghan, Christiansen, & Chater, 2004), among others. The usual procedure in such experiments is the same artificial language paradigm we have been discussing. The great virtue of an artificial language is that one has far more control over the stimulus.

Research from Gomez (2002, 2006) and Onnis et al. (2004) is particularly informative regarding the role of statistical structure in language learning and parsing. Gomez (2002) and Onnis et al. (2004) both looked at pattern learning for phrase-like structures. The design of their language followed an A-\(X\)-B pattern, in which B was highly predicted by A with an intervening variable \(X\). Both infants and adults in their work succeeded or failed to learn the language depending upon the variability of \(X\). Under very low or high variability of the syllable \(X\), the language is learned. Under "medium" variability, the language is not learned, suggesting that when A had “medium”
correlation with $X$, participants focused there. When the relationship to $X$ carried little information, they switched to focus upon the non-adjacent relationship. This directly implicates the stochastic patterns of the language in triggering learning. Not only is variability a possible way to learn the language (and hopefully the only one so that there are no confounds) as in the Saffran-type experiments, variability is directly changing language learning behavior.

We also have evidence of linguistic behavior mimicking probability distributions in a fine-grained manner. Vouloumanos (2008) trained adults on object labels. Some labels went to objects in a one-to-one relationship, while other labels were used for multiple objects. This creates a probability distribution based on the chance of a label referring to a given object. Adults not only learned deterministic and high probability pairings; they also discriminated among rather infrequent mappings, such as finding a 20% pairing more likely than a 10% pairing. Clayards, Tanenhaus, Aslin, and Jacobs (2008) also found evidence that people were sensitive to an entire probability distribution. They created two probability distributions of voice onset times in bilabial stops, each with the same mean but with either wide or low variance from the mean. Participants displayed confidence levels that were steeper or shallower depending on the size of the variance; i.e., wider statistical variance had shallower confidence curves.

1.3.3. But what is statistical learning exactly?

*How to Calculate.* If statistical distributions are potentially useful across phonological, lexical, syntactic, and semantic domains, as the previous research attempts to show, then we certainly need to understand the mechanism in detail. This is one area where there is no consensus in the literature. Saffran and colleagues, as well as many
commentators and fellow researchers, describe their result as a demonstration that adult speakers can find dips in transitional probabilities in speech and will posit word boundaries at those locations, but this is not the only legitimate explanation. One could also take the opposite perspective and say that the speakers were discovering clusters of highly predictive syllables (Swingley, 2005). The result is the same, but the mechanism is different, and this could be quite relevant in later research.

There are also multiple statistical calculations that could explain the results. Aslin, Saffran, and Newport (1999) point out that forward transitional probability, \( p(s_2|s_1) \), backward transitional probability, \( p(s_1|s_2) \), and information theoretic measures such as conditional entropy or mutual information would all distinguish words from part-words. However, importantly, simple co-occurrence is not sufficient. Aslin et al. (1999) modified the language in order for part-words and words to occur an equivalent number of times, setting raw frequencies to be equal, while transitional probabilities remained higher for words. Speakers still located the words within the speech stream.

Perruchet and Desaulty (2008) provide evidence that backwards transitional probability is the correct measure, not forwards transitional probability, while Onnis (submitted) found language specificity regarding this distinction. Specifically, English-speaking participants learned an artificial language based upon backward probability while Korean participants based their decisions off of forward probability. Finally, Perruchet and Peereman (2004) discovered the best statistical match for judgments from French speakers about phonotactic distributions was the r-phi measure, which is roughly a bi-directional version of forward and backward transitional probability.
From this, we can conclude at a minimum that the precise statistical calculation is not settled. Moreover, there may be no single calculation for adult speakers for all tasks in all languages. Instead, a more general learning ability must be tuned to fit the task that the speaker faces in their linguistic environment. Perhaps work with infants, instead of adults, could address the beginnings of these abilities.²

What to Calculate. One reason that statistical learning was not a prominent topic for linguists for many years is that unconstrained statistical calculations would be of little use. We have assumed in our discussion of Saffran et al. that the listeners were calculating the statistical structure of syllables, but why not segments? Or syllable codas? Or the phonetic burst frequencies on every fourth occurring voiceless stop when spoken by a female speaker on Tuesdays? Yang (2004) argues that Universal Grammar is necessary to constrain statistical calculations (and that statistical calculations are needed to correctly parameterize Universal Grammar). Others have attempted to use a bootstrapping method to solve such issues. For instance, Maye et al. (2002) derive phonemic categories from phonetic distributions. In turn, this would allow an infant to focus upon some unit like a phoneme or segment for later learning. Still other researchers have proposed different mental mechanisms for different tasks (Endress & Mehler, 2009; Endress, Nespor, & Mehler, 2009).

One place this issue has received significant attention is the difference between calculating probabilities for adjacent and non-adjacent items. If we only calculate statistics for adjacent units, this highly constrains the statistical learning process. However, on the surface, language is filled with non-adjacent dependencies, such as

² We address this again in Section 8.2.1 and suggest a measure that is independently motivated.
subject-verb agreement and certain morphological forms, so a purely adjacent learning could be insufficient to learn linguistic structure.

Word segmentation and basic grammars that depend upon adjacent relationships are generally learned rather easily. However, there seems to be much greater difficulty when the patterns are based in units separated by one or two intervening items. In a word segmentation task, adults appear to be able to learn patterns where there are high probabilities between nonadjacent consonants or vowels (patterns such as bXdYtZ or XaYiZe where uppercase letters represent variables and lowercase letters represent phonetic segments), but not when the high probabilities are between nonadjacent syllables (patterns such as baXte; Newport & Aslin, 2004).

However, Onnis et al. (2004) showed that nonadjacent syllabic patterns can be learned when the variability in the "middle" syllable is sufficiently high. They created two languages, one where 3 syllables instantiated the middle syllable and one where 24 syllables instantiated the middle syllable. Adults failed to find words under the former condition but succeeded in the latter. The language used in Newport and Aslin (2004) more closely matched Onnis at al.'s low variability condition. Based upon these results and many studies in her own lab (Gomez, 2002), Gomez (2006) suggests that speakers preferentially extract information from adjacent dependencies and then, only when they fail to find structure in their preferred search pattern, do they switch to nonadjacent dependencies. She terms this dynamically guided learning.

For this dissertation, we do not need to resolve the issue of adjacency versus non-adjacency or what the units of statistical calculation are, but there are a few items that are worth noting. First, adults appear to be able to make judgments in word segmentation
tasks based upon both syllabic probabilities (Johnson & Jusczyk, 2001; Newport & Aslin, 2004; Saffran, Newport, & Aslin, 1996) and segmental probabilities (Newport & Aslin, 2004), particularly consonants (Bonatti, Pena, Nesport, & Mehler, 2005). Secondly, participants are constantly seeking for structure in the input. Gomez’ dynamically guided learning proposes that the participant will be pre-disposed to search for structure in one location first, and then, when structure is not found, switch to another. This is implicit in the previous discussion of participants first segmenting speech based upon immediately available cues and only switching to a different one later if the first is insufficient.

This search for structure need not be entirely sequential. Indeed, we have evidence that people can employ several hypotheses at once. In a visual categorization task, participants classified shapes while wearing an eye-tracking device. On the whole, participants looked at a large number of dimensions during the early portions of the training and only collapsed attention to the relevant dimensions, ignoring irrelevant ones, after the problem had been solved. Once the relevant dimension was known, participants focused selective attention there. This is to be opposed to immediately focusing on one dimension of the problem that is assumed to be relevant, trying it out to see if any progress is made, and then moving to the next dimension in a series of steps until the problem is solved (Rehder and Hoffman, 2005). In an auditory task with sound sequences that included sporadic unexpected tones, participants were able to follow both adjacent and non-adjacent patterns simultaneously, demonstrated by a Mismatched Negativity (EEG) response when either pattern was violated (Horváth, Czigler, Sussman, & Winkler, 2001).
The overall picture then is that we are able to track some number of cues simultaneously searching for information. However, the incremental and immediate nature of human processing can lead us down a “garden path” (Gebhart et al., 2009). If we locate sufficient structure, we could remain with this conclusion even if it is not the ideal conclusion. (In computational learning theory, we might consider this a local minimum.) If there is sufficient evidence over time that this was a mistake, we can get off the garden path and discover a more complete structure (Gomez, 2006; Onnis et al., 2004). Unfortunately, we do not know the exact dimensions of either gardenpathing or recovery.³

1. 4. The Experience of Uncertainty

At this point, it is worthwhile to pause and assess our progress so far. We have discussed the apparent importance of statistical learning to our understanding of both language acquisition and language processing. We have also reviewed work on both the units upon which statistical calculations might operate and what those calculations might be. One thing we have not addressed is what it might be like to experience statistical calculations. This may seem an odd issue to be concerned with. Statistical calculation is clearly subconscious. No one believes an 8-month-old consciously counts syllables and tries to keep a mental tab. The calculations necessary to accomplish a word segmentation experiment using transitional probability are formidable. Aslin et al. (1999) point out that their word segmentation task involves "the on-line (running) computation of 20 different

³ This is not to say there are not proposals of various sorts from dynamic system theory to Bayesian rational analysis to information-theoretic neural computation and beyond. In Section 8.2.1, we will make some tentative suggestions that “surprisal” in the sense of Levy (2008) may be the ideal measure within the context and concerns of this dissertation.
conditional probabilities, each over 45 to 90 occurrences of the component syllables and 9 to 90 occurrences of syllable pairs, during a 3-min learning period" (p. 323, fn. 4).

If transitional probabilities are calculated, they must be a general feature of neural computation or other basic mental structures, such as memory formation and retrieval. For instance, Denham and Winkler (2006) argue that prediction is fundamental to any auditory streaming (Bregman, 1990), without which all auditory experience would be confused and fragmentary.

Simply because statistical calculations cannot be consciously observed, however, does not mean they are not subjectively experienced. A tip-of-the-tongue state is highly salient subjectively, even though we cannot consciously observe what is happening in our mind. To attempt to discover what it is like to experience a statistical calculation, let us revisit the Saffran, Newport, & Aslin (1996) experiment once again (henceforth: Saffran et al., 1996, as we will not be concerned further with the work with infants). The typical explanation of what participants are doing here is calculating the forward transitional probability between syllables – the probability of syllable j given syllable i. We can call these the SylProbs. The transitional probabilities are higher within words than across a word boundary, so there will be a dip in SylProbs at each word boundary.

The essential psychological notion behind forward transitional probabilities is that the listener is hearing one piece of language and making a prediction for the next. They might be predicting the next phoneme, syllable, syntactic structure, semantic role, etc., but they are always making predictions for what will happen next. In the artificial language of the Saffran et al. experiment, their predictions for the second and third syllables of any trisyllabic word will be relatively good, since they regularly follow with
high transitional probability. However, when the third syllable is heard, suddenly the participant does not have a clear idea what syllable will come next. This is because they are at a word boundary, though they may not be aware of the reason. Locally, they simply cannot predict with much success what the next syllable will be. Uncertainty then rises at the end of each word in the Saffran et al. experiment.

While this is oversimplification, the underlying experience of knowing the statistics of the Saffran et al. language is that one is (relatively) certain after hearing the first syllable, (relatively) certain after hearing the second syllable, but (relatively) uncertain after hearing the third syllable. The major proposal here is that uncertainty is experienced as hesitancy, i.e., as an increase in duration. We will review the logic for this in the following.

1.4.1 Subjective time perception and predictability

The average person is aware that subjective time does not equal physical time. In a scene from the TV show Star Trek: The Next Generation, there is a character, Commander Data, who is an android and therefore only experiences time using an embedded physical clock that is assumed to be invariable in the way real-life atomic clocks are invariable and absolute at least on a human time-scale. In one scene, Data’s supervisor enters Data’s private quarters to find him boiling water. Data states:

I have been testing the aphorism, “A watched pot never boils.” I have boiled the same amount of water in this kettle sixty-two times. In some cases I have ignored the kettle; in others, I have watched it intently. In every instance, the water reaches its boiling point in precisely 51.7 seconds. It appears I am not capable of perceiving time any differently than my internal chronometer. (Star Trek: The Next Generation, 1993)

---

4 In the experiments for this dissertation, there is a probability of 1 within words.
Colloquialisms about the subjective perception of time by non-androids are common: “A watched pot never boils;” “time flies when you are having fun;” etc. We have other anecdotal experiences of time perception being variable. When we are uncertain where we are going, such as going out on a hike to a new area, it typically seems to take longer to hike out than it does to hike back, when we now know the way, regardless of whether it actually takes longer. People have reported time moving slowly in traumatic situations or at points of heightened attention, such as at a pivotal moment of an athletic event.

Beyond idioms and anecdotes, there is a large body of psychological work over 4 decades unraveling the subjective perception of time. In a touchstone paper for this subfield, Hicks, Miller, & Kinsbourne (1976) showed that when attention was increased on a task, perceived durations increased. Similarly, attention away from an event would decrease the perceived duration. Many studies since have focused upon the relationship between attention, subjective time, and information processing. (Irvy & Schlerf, 2008, provide a solid review of this literature.)

In one important study, Tse, Intriligator, Rivest, & Cavanagh (2004) used a number of subjective time measurements with a series of stimuli that included one low-probability “oddball” stimulus after several predictable stimuli. Participants judge oddballs as lasting longer even when they are presented for the same amount of time as the predictable stimuli. They summarize their findings thus:

We agree with the traditional view (e.g., Creelman, 1962; Thomas & Weaver, 1975; Treisman, 1963) of time perception, according to which perceived duration is a function of the amount of information processed per unit of objective time. We also accept the standard view that attention can influence the perception of duration (Tse et al., 2004, p. 1184).
They go on to suggest that the reason for the change in perception is that increased attention allows more information processing to occur. Under this theory then, time is subjectively measured by the amount of information processed. With increased attention, more information processing can occur over the same amount of objective time than can occur over that same time with decreased attention. The result is that attention modulates information processing which in turn affects the perception of time.

This overall notion is quite compatible with the psycholinguistic tradition. We typically assume that increased processing will take an increased amount of time. This is one of the bases for using Reaction Time to explore cognitive behavior. Employing unexpected syntactic structures increases reaction time; retrieving unexpected words increases reaction time; etc. When reaction times are longer in an experiment, we assume that more (or at least more difficult) processing occurred. The usefulness of this is tied to how good our theoretical models are at describing the underlying mental processes. If the models are wholly inaccurate, then we simply know that some task took longer, but not why. If our models are approximately correct, then the difference in time provides evidence that one path was followed, not another. Tse et al. (2004) are adding the idea that the perception of time will not directly reflect actual computing time, but be modulated by resources devoted to processing the information. This is potentially important for a large range of linguistic issues. For instance, a syntactic processing experiment may be interested in the duration of a pause at the end of a phrase and its impact upon parsing. If the subjective time literature is accurate, the perception of that

---

5 In the terms of rate distortion theory (Shannon, 1948), it is as if attention can modify channel capacity dynamically.
physical duration could change depending upon the predictability of the linguistic material.

Pariyadath and Eagleman (2007) have been able to take these notions further in a couple of ways. First, in further work with a series of stimuli, they crossed a condition that should increase attention, using emotionally charged but predictable items, with a condition that used an emotionally neutral but unpredictable oddball stimulus. It was only the latter unpredictable stimulus that triggered perceptions of increased duration, suggesting it is predictability, not attention that modulates subjective time perception. Secondly, they compared the perceived duration of the first stimulus in a series to the perceived duration of following stimuli in the series. In series in which the first stimuli predicted later stimuli, the first was judged as having a longer duration than following ones. However, when the series was composed of a random sequence, so that following stimuli could not be predicted by the first, the judged durations were equivalent. This suggests again that predictability decreases subjective duration, while unpredictability increases it.6

Unfortunately, we cannot map the results of Pariyadath and Eagleman (2007) directly to the word segmentation experiments, as the tasks are slightly different. In the oddball series, participants are seeing a series of predicted stimuli and suddenly an unexpected one arrives. The unpredictable oddball is perceived as longer in duration. With the word segmentation experiments, the unpredictability is not that something surprising suddenly appeared when participants thought they knew what would happen

---

6 One limitation to this work is that it all uses a variety of retrospective, though immediate, time judgments. There is a possibility that a retrospective judgment, even if given immediately after the stimulus, does not directly reflect the perception of time during the stimulus.
next. Instead, if they are calculating transitional probabilities, they know at the end of each word that the next syllable is uncertain. When the next word arrives, uncertainty actually disappears. Therefore, in the word segmentation experiments, the temporal locus of uncertainty is actually on the third syllable, not in the first syllable of the next word. If the perceived increase in duration occurs at the point of unpredictability, then that extra lengthening would be heard either on the third syllable itself or as a possible pause after the syllable. Putting all of this together, we can posit the Information/Duration Hypothesis for linguistic processing.

**The Information/Duration Hypothesis**: An increase in uncertainty during linguistic processing is perceived as an increase in duration. A decrease in uncertainty is perceived as a decrease in duration.

Note that, in this formulation, it is a change in uncertainty that is perceived as a change in duration, not a constant level of uncertainty, though that possibility should be explored. Also, the Information/Duration Hypothesis (IDH) is only formulated in one direction: Uncertainty increases perceived duration, but greater duration may not increase uncertainty.

---

7 In the segmentation experiments to be discussed in Chapters 4-6, the researcher would often chat informally with the participant afterwards about their experience. Some participants would report that they heard breaks between the words after a while, breaks that do not physically exist. Unfortunately, these reports were not catalogued and cannot be compared with actual behavior on the tests. In the instances that were examined at the time, there was often little correspondence between the participant’s self-rating of success and their test score.

8 This notion begs an interpretation in terms of information theory and entropy. This is beyond the scope of the current work but is a promising topic for future work.
1.5 Prosodic Boundaries, the IDH, and Word Segmentation

If the IDH were to be correct, it could have broad implications for linguistic processing as it suggests that the perception of duration is mediated by the expectations of the linguistic material. As expectations are dynamically constructed, duration perception would change dynamically during speech perception as well. In the artificial language word segmentation experiments that are the focus of this dissertation, the IDH would trigger the perception of a slight lengthening in duration at the end of each word in the speech stream – only after the statistical structure has emerged, however. It is the transitional probabilities that create the difference in uncertainty that occurs at the end of each word. Recall that Pariyadath and Eagleman (2007) saw no changes in duration perception with a random series of stimuli.

Let us apply this to the Saffran et al. (1996) stimuli. The words in that artificial language were: babupu, bupada, dutaba, patubi, pidabu, and tutibu. If we concatenate them together randomly, the experience orthographically would be something like (4).

(4) bupadatutibabupupatubibupadabupupidabututibu…

After calculating transitional probabilities, however, and applying the IDH, there will be the percept of a slight lengthening at the end of each word; i.e., on the third syllable since there is uncertainty at that point as to what comes next. We can represent this with the [:] marker.


We mentioned briefly in Section 1.3.1 that word final vowel lengthening is a documented cue for word segmentation (Salverda, Dahan, & McQueen, 2007). The
Saffran et al. (1996) study that we have concentrated on so much has a second experiment in which the final vowels of each word in the artificial language were lengthened by 100 ms. The non-lengthened syllable durations were approximately 278 ms so this corresponds to a 36% lengthening. In this condition, participants selected words correctly 80% of the time on average compared to 65% when there was no lengthening, a significant difference. This provides further evidence that lengthening is a cue and that, if there is lengthening due to the IDH, it does not correspond to a 35% increase.

1.5.1 Segmentation through prosody

Through much of this chapter we have spoken of transitional probabilities as marking word boundaries. The IDH raises the possibility that transitional probabilities can mark prosodic boundaries, as the ends of prosodic units in English are typically lengthened. There is a possibility that what is called final word lengthening is in fact prosodic lengthening which is aligned with the word. Salverda et al. (2007) compare monosyllabic words such as ham to the same sequence of phonemes in the first syllable of the word hamster. In an eye-tracking study, they discovered that the longer the syllable ham of hamster is in duration, the more participants look at a picture of a ham and not a hamster. They argue that this is because a prosodic boundary typically follows content words, so that a very long ham is seen as taking up an entire prosodic unit and must therefore be the end of a word instead of the first part of a disyllabic word.

Indeed, prosody is intimately tied with word segmentation in various ways. Speech is lengthened at the end of a prosodic unit and a word. We have already discussed how the prosodic bias of English towards trochaic strong-weak syllable pairs is a stronger
cue for word segmentation than statistical patterns (Johnson & Jusczyk, 2001; Johnson & Seidl, 2009). Cutler and Norris (1988) formalize this as the Metrical Segmentation Strategy in which listeners use their knowledge of metrical phonology to segment speech, and metrical patterns are widely documented to affect speech segmentation (Cairns, Shillcock, Chater, & Levy, 1997; Curtin, Mintz, & Christiansen, 2005; Jusczyk, 2001; Murty, Otake, & Cutler, 2007). Moreover, segmentation tasks that can be difficult are eased when a prosodic boundary is explicitly present. English-acquiring infants have a difficult time segmenting vowel-initial words but can do so when an explicit prosodic cue is placed before the initial vowel (Seidl & Johnson, 2008).

It is tempting to suggest that all word segmentation is phonologically mediated so that so-called word segmentation cues are prosodic cues that are then used for segmentation. We need not go so far. The fact that morphological words do not always align perfectly with phonological words, such as in the case of clitics, should hold us back from this conclusion. However, it is clear that prosody does influence word segmentation such that we should certainly look for any prosodic cues when interpreting the results of a word segmentation experiment. Moreover, we have previously discussed that language processing is immediate and incremental (Section 1.3.1). This naturally includes phonological processing. The participant will attempt to map their native prosody to a language stimulus, helpful or not. The most common setting for this in a natural context is with a second language. American English speakers, for instance, will reinterpret stress-less Japanese words as containing stress.
1.5.2 Rhythm and word segmentation

The idea that prosody affects word segmentation is no revelation. Studies of word segmentation generally divide into research on distributional patterns, prosodic patterns, and the relation between the two. In our overview of the Saffran et al. (1996) experiment, we discussed how they attempted to control for prosodic cues by having all syllables be the same duration, intensity, and pitch. They also controlled for co-articulation so that the segmental phonetics would not betray the metrical structure. However, it may not be possible to truly remove prosodic cues from a stimulus. If one makes all durations identical, so as not to trigger stress perception, one has given the language an explicit rhythmic structure. Every syllable onset is isochronous. Since all words have the same number of syllables, every word onset is also isochronous. This essentially puts a rhythmic beat upon each word onset. If the beat can be found, then one can segment the entire language without calculating a single statistic. We turn to a detailed exploration of the perception of time, prosody, and the implications for word segmentation in the next chapter.
2.1. Time in Language Research

2.1.1. Stress, syllable, and mora timing.

Historically, there have been two primary areas of research into speech rhythm. The first might be labeled unit-timing and the second the study of stress, particularly as seen through the lens of metrical phonology. Focusing first on unit-timing, Pike (1945) distinguished two classes of languages: stress-timed and syllable-timed. In stress-timed languages, inter-stress intervals were argued to be isochronous, while in syllable-timed languages, inter-syllabic intervals were isochronous. Typical stress-timed languages include the Germanic languages such as English, Dutch, and German, while French, Spanish, Korean, and Mandarin are syllable-timed. The difference between these two rhythm classes lies largely in which phonological unit "drives" timing, as stress-timed languages are universally composed of syllables and many syllable-timed languages have stress. This notion of isochrony can be seen in (6) and (7).

(6) Stress-Timed Data (Speech in orthography; stress marked with CAPS; time noted in seconds.milliseconds format):

<table>
<thead>
<tr>
<th>Syllables: Alice</th>
<th>RITE</th>
<th>WAL</th>
<th>ker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Points:</td>
<td>0.000</td>
<td>0.200</td>
<td>0.300</td>
</tr>
<tr>
<td>Syl. Interval:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syl. Duration (ms):</td>
<td>200</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Stress Interval:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stress Dur. (ms):</td>
<td>300</td>
<td>300</td>
<td></td>
</tr>
</tbody>
</table>

(7) Syllable-Timed Data

<table>
<thead>
<tr>
<th>Syllables: Marie</th>
<th>Gene</th>
<th>VIEVE</th>
<th>Du</th>
<th>PONT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Points:</td>
<td>0.000</td>
<td>0.200</td>
<td>0.400</td>
<td>0.600</td>
</tr>
<tr>
<td>Syl. Interval:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syl. Duration (ms):</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Stress Interval:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stress Dur. (ms):</td>
<td>600</td>
<td>400</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Looking first at (6), the time between each syllable varies from 200 ms to 100 ms to 150 ms. However, the time between each stressed syllable is a constant 300 ms. For (7) on the other hand, the interval between each syllable is a constant 200 ms, while the time between stressed syllables changes from 600 ms to 400 ms due to different numbers of syllables between stresses.

While the concept of rhythm classes has persisted until the present day, the actual basis of these divisions has become quite debatable. Bolinger (1965) demonstrated that the time between stressed syllables changed depending upon the type of syllables involved; Wenk and Wiolland (1982) did not find isochronous syllables in French; Roach (1982) compared three ostensibly stress-timed languages against three syllable-timed languages and found comparable syllable durations in all. Lehiste (1973, 1977) argued that stress and syllable isochrony were perceptual, not physical, and that the expectation of isochrony allowed for linguistic information to be embedded precisely in the violation of this expectation. However, what is the perception of isochrony based upon?

Several recent attempts have found fairly reliable acoustic cues to rhythm class based upon overall consonant, vowel, or syllable durations. Based upon the notion that stress-timed languages and syllable-timed languages frequently have different sorts of syllable structures and phonological patterns (relatively complex consonant clusters and vowel reduction in stress-timed languages as opposed to fewer consonant clusters and a lack of vowel reduction in syllable-timed languages), Ramus, Nespor, and Mehler (1999) examined the standard deviation of vocalic and consonantal intervals (labeled $\Delta V$ and $\Delta C$ respectively), as well as the overall proportion of vocalic intervals in a sentence ($%V$). They found that such metrics could distinguish languages by their theoretical rhythmic
classes. For instance, in a chart sorted by the %V metric, the stress-timed Germanic languages cluster together, followed by a group of syllable-timed Romance languages, with mora-timed Japanese as the last member of the chart. The ΔC, ΔV, and %V metrics have been enhanced by later research. One critical development is the use of a normalization factor to compensate for differences in speech rate (Deterling, 2006; Low, Grabe, & Nolan, 2000; White & Mattys, 2007).

Ramus et al. (1999) and White and Mattys (2007) explicitly argue that their results suggest a different explanation for unit-timing than the traditional "top-down" proposal. White and Mattys state, "...[D]ifferences in rhythm scores between languages arise from language-specific variations in syllable construction and distribution, as well as from segmental timing processes. The rhythmic distinctions suggested by these metrics are therefore emergent phenomena" (2007, p. 518). In other words, against Pike (1945), languages do not have overall rhythms into which segments are placed and manipulated. Instead, so-called syllable-timed languages sound syllable-timed simply because they have fewer consonant clusters allowing for more even syllable durations. Before we can assess this claim, we first must take a look at the other main approach to rhythm in languages: stress and metrical phonology.

2.1.2 Metrical phonology

I will use the term 'stress' in the sense of metrical phonology (Liberman, 1975; Liberman & Prince, 1977; Prince, 1983). Hayes (1995) defines stress in this sense as "the linguistic manifestation of rhythmic structure. That is, in stress languages, every utterance has a rhythmic structure which serves as an organizing framework for that utterance's phonological and phonetic realization" (p. 8). Liberman and Prince (1977) use metrical
grids to understand stress, with a grid defined as "hierarchies of intersecting periodicities" (p. 313). The syllable, for English, is the most basic rhythmic level with each syllable receiving a place in the metrical grid. The next level up marks the foot level. Example (8) displays a grid for the text from (6) in the bracketed grid notation of Hayes (1995).

(8) Alice Rita Walker
   |x   |x   |x   |
   |x x |x x |x x |

On the lowest level of (8), each syllable is marked with an $x$ to give it a “beat”. Each of these words is stressed on the first syllable, and so, on the second level, another $x$ is marked. Borrowing terms from poetic meter, the lower level is the syllable level and the upper level is the foot level. As discussed in Chapter 1, approximately 80% of content words in English follow this pattern of a stronger beat on the first syllable of a foot and a weaker beat on the second. One can see that this is the default metrical pattern of English through the way it adopts loan words. For example, Japanese is a language without stress, but when English borrows a Japanese word, the trochaic pattern is imposed. The company name *Mitsubishi* has no stress in Japanese, but is pronounced as in (9) by English speakers.

(9) Mit su bi shi
    |x   |x   |
    |x x |x x |

When pronouncing a word such as *Mitsubishi*, however, both stressed syllables are not equally strong. Instead, the next to last syllable *bi* is pronounced strongest of all: *MitsuBiShi*. This is the layer of the prosodic word, and the general pattern is to pronounce the final stress as the strongest syllable, as seen in (10). Again, this pattern will be applied
to novel loan words, so it cannot arise either from memorization or from hearing it. It is part of the phonological grammar of English.

(10) Mit su bi shi
Pros. Word: | x | x | x |
Foot: | x | x | x |
Syllable: | x | x | x | x |

One of the great linguistic contributions of this work is that, predominantly through a typology of feet and alignment, one can account for the stress patterns of most languages of the world. For instance, feet can be composed either of two syllables, as in English, or two mora, as in Hawaiian. A mora is a short vowel such that a syllable with a single short vowel is made of one mora and a syllable with a long vowel is made of two mora. The stress pattern of Hawaiian is roughly the same as English except that the feet are composed of moraic trochees, instead of syllabic trochees. Other languages may employ a foot which is weak – strong, instead of English’s strong-weak. Such feet are termed iambs, another term borrowed from poetic analysis.

While the most common pattern of English words is trochaic, and that is the pattern applied to novel forms, iambic feet are possible. Many noun / verb pairs are distinguished in English only by their stress pattern.

| Table 1. Words Distinguished by Stress in English (stress indicated with CAPS) |
|---------------------------------|---------------------------------|
| Word 1                          | Word 2                          |
| INsult (noun)                   | inSULT (verb)                   |
| PERvert (noun)                  | perVERT (verb)                  |
| PERmit (noun)                   | perMIT (verb)                   |
| PROduce (noun)                  | proDUCE (verb)                  |

Table 1 is of course not exhaustive. Hayes argues that stress languages around the world can largely be accounted for by a system of binary feet of three types: syllabic trochees, moraic trochees, and iambs. If stress languages are made up of three types of feet, all of
which are binary, what do we do with one-syllable words? Section 1.5.1 mentioned the case of *ham* versus *hamster* in a study by Salverda et al (2007). We saw there that the longer the syllable *ham* was, the more participants thought it was the monosyllabic *ham*, instead of the first part of the disyllabic *hamster*. While we avoided the terms then, because they had not been introduced, essentially, the monosyllabic *ham* is lengthened to fill a metrical foot.

In Hayes’ (1995) metrical theory, a foot with a single syllable (within a syllabic trochee system) is termed a degenerate foot. Some languages forbid all degenerate feet. Hawaiian, for instance, which uses a moraic trochee system, can only have a word that is at least two mora in length. English does allow degenerate feet, but only if they are also prosodic words. This has the effect that a degenerate foot must be a word in English and cannot be part of a polysyllabic word.9

One final term we must discuss is that of “extrametricality”. A syllable is said to be extrametrical if it remains unparsed; i.e., not part of any foot. For a variety of reasons, Hayes (1995) argued that feet are exclusively binary across languages. Many English nouns, in which the last stress appears to be the antepenultimate syllable, i.e., three syllables before the edge, seemingly contradict this. Consider the word *serendipity* [səˈrɛn.dɪ.pɪ.tɪ]. If English has only binary feet, then there should also be a stress on the final 5th syllable of *serendipity*, since there are clear stresses on the 1st and 3rd syllables. However, speakers do not hear the last syllable as stressed; it often has little in the way of acoustic markers of stress; and, moreover, the /t/ is flapped to a [ɾ], which happens precisely when the syllable that it initiates is unstressed (compare *atomic* [ə.təˈmɪtɪc] and

---

9 These terms will be used extensively during the analysis of the results in which we see how the participant is mapping their phonology onto the artificial language.
To handle such cases (and for many other reasons as well) Hayes (1995) allowed for some syllables to be extrametrical – not part of a foot at all.

Three possible metrical parses of *serendipity* are provided in (11), (12), and (13). In (11) the final syllable is given its own foot. However, that implies the final syllable is stressed and even a prosodic word, if we follow the rules of degenerate feet sketched above. Since there is no evidence of stress on that syllable, (11) can be rejected. In (12) the final foot is extended to be trisyllabic. This provides the correct stress assignments, but violates Hayes’ goal of only having binary feet. Finally, (13) allows the final syllable to be extrametrical – or not in a foot at all. Notice in the transcription that the syllable *ty* is not part of a foot and is marked with < >. While this is an oversimplification, the choice between (12) and (13) is in part a matter of whether ternary feet (feet with three units) or extrametricality is the simpler theory.

(11) se  ren  di  pi  ty
     |    | x    | x    |
     | x    | x    |
     | x    | x    | x    |

(12) se  ren  di  pi  ty
     |    | x    |
     | x    | x    |
     | x    | x    | x    |

(13) se  ren  di  pi  ty
     |    | x    |
     | x    | x    |
     | x    | x    | x    |<x>

A virtue of metrical phonology, besides its admirable coverage of many of the world’s stress patterns, was that it gave a natural explanation to a phenomenon that had long been known but poorly motivated, namely the resolution of stress clash. If two
strong beats—two stressed syllables—become adjacent in a spoken phrase, there is a strong tendency among English speakers to change the location of one of the stresses such that an unstressed syllable intervenes. A classic example from Prince (1983, p. 21; Hayes’ brackets have been added) illustrates this:

(14) a) \[\underline{x} \underline{x} \underline{x} \underline{x}\] b) \[\underline{x} \underline{x} \underline{x} \underline{x}\] c) \[\underline{x} \underline{x} \underline{x} \underline{x}\] d) \[\underline{x} \underline{x} \underline{x} \underline{x}\]

Examples 14a) and 14b) provide the stress pattern of each word when spoken alone. Stress is on the second syllable for *fourteen* and on the first syllable for *women*. When the two words are combined to make a phrase as in 14c), however, a stress clash occurs. Two stressed syllables are now adjacent and there is no intervening weak syllable. 14c) is usually rejected then, and the rhythm is fixed by moving the stress of *fourteen* to the first syllable *four*, as seen in 14d).

*Metrical phonology and unit-timing.* One can see an obvious connection between metrical phonology and stress-timing. Liberman and Prince (1977) state at one point that metrical grids are a formalization of the traditional idea of stress-timing. Hayes (1995) suggests isochrony for each foot, which would also generate stress-timing. However, there are some discrepancies between metrical phonology and unit-timing as well. It is not clear how to handle syllable-timing, for instance. If isochrony is driven from foot to
foot, it suggests that syllable-timing is derivative from the feet, which opposes the idea that syllables are the primary driver determining timing in syllable-timed languages, not stress. Also, how do we handle languages that are said to be syllable-timed, such as Korean, that do not have stress at all? Moreover, metrical phonology is at root a structural analysis of strong and weak patterns. To map it directly to timing, we would need an extra set of processes converting the grid into a time series.

The tradition of metrical phonology is also at odds with the work of Ramus et al. (1999) and White and Mattys (2007). Metrical phonology places almost all of speech rhythm into a structure with few references to syllable content.\textsuperscript{10} The structure is driving the rhythm. However, White and Mattys (2007) concluded that rhythmic differences were emergent phenomena reflecting patterns of consonant clusters and vowel reduction. They then allow little room for structure in unit-timing.

A weakness of the “emergent” approach to unit-timing is that, even if natural speech were to lack predictable isochronous intervals, we seem to be able to create such intervals in the right conditions, such as when a metronome induces a clear rhythm. Port (2003) asked participants to repeat a phrase several times with a metronome. Under such conditions, participants create isochrony by aligning stressed syllables in largely similar ways from speaker to speaker. This is clearly not a natural task, but if there were no underlying periodic tendencies within speech, it is not clear how participants were able to do this with little training and in similar ways. More natural situations in which we can use language in a highly rhythmic way include music, chant, and some poetry. There

\textsuperscript{10} Metrical phonology does include notions of syllable weight, but they typically exclude consonant clusters in the onset, which are highly relevant in the measures of Ramus et al. (1999).
may be a way to unite these disparate senses of timing using temporal oscillations. We turn now to that topic.

2.2. Temporal Periods Understood as Oscillators

One of the more productive explorations of rhythm in psychology over the last three decades is that of dynamic attending from Jones and colleagues (Jones, 1976; Jones & Boltz, 1989; Large & Jones, 1999). In this theory of rhythm perception, humans have internal rhythms that can become entrained with an external stimulus. Each individual has a unique referent period, which is most easily revealed by tapping a finger naturally without any external stimulus. This referent period is unique for each person, but it changes throughout one's lifetime. Adults' referent period centers around 600 ms, while that of a child of 4 to 6 centers around 400 ms (Drake, Jones, & Baruch, 2000). This referent period can become entrained with an external stimulus, providing the referent level that corresponds to the level of the tactus, the most basic level of rhythm where attention seems to naturally fall and to which the foot typically taps.

This is modeled formally by a series of dynamic coupled oscillators (Large & Jones, 1999; Large, 2008). To define these terms, a pendulum on a clock is a simple oscillator that swings between two states over time. An analog clock face can also be thought of as an oscillator in which the hands move through space with a certain period. A pendulum, however, is not coupled with its environment (discounting friction). It swings through its cycle based upon its own characteristics. If it were to alter its period based upon another pendulum, then it could be considered coupled. When a drummer begins a periodic rhythm and a guitarist then joins in, strumming with the same period and on the same beat, then the guitarist’s oscillatory motions with the hands are coupled
to the oscillations of the drummer’s arms. Naturally, if each oscillators can dynamically adjust its behavior to remain synchronized with the other, then they are dynamically coupled oscillators. Musicians playing together are dynamically coupled.

In Large and Jones’ (1999) theory, there are three critical variables to the model. First, the period of the oscillator describes the time of the entire cycle. The phase variable marks the state of the cycle at any given time \( t \). If we are using the analog of a clock face to understand oscillators, the period describes the amount of time it takes for a hand to make a complete cycle of the clock face. The phase angle variable states where on the clock face a hand is at any given point. Phase angles are often stated in degrees, radians, or simple phase-units. Using the degree notation, when the hand is pointing to 12 o’clock, that would be a phase angle of 0°. Six o’clock would correspond to a phase of 180°. If a process oscillated between two states, that binary oscillation could be stated as moving between phase(0°) and phase (180°). We could conceive of a triadic process on the clock face as well. Here, 4 o’clock would have a phase angle of (120°) and 8 o’clock would have a phase angle of (240°). If one translated the phase angle from degrees to units based upon the oscillator’s period, then a binary oscillation would have two primary phase angles of (0) and (1/2), while a ternary oscillation would have three primary phase angles of (0), (1/3), and (2/3).

The third variable of the Large & Jones model is an attentional pulse (making this clearly rhythmic attending) that occurs at phase (0) of the cycle. This attentional pulse is modeled with a von Mises probability distribution, a common distribution for cyclic mechanisms. This distribution can become sharpened, creating a greater and more precise expectation at phase (0), by increasing the focus variable. This expectation increases as
the external stimulus synchronizes (is coupled) with the internal oscillator. Low synchronization creates a shallow, distributed expectation, while high synchronization creates a high and narrow expectation. This matches intuitively with our notions of the effects of rhythmic structure in a sound. The white noise-like characteristics of falling rain create no expectation of when the next event will occur, while extremely structured music lets virtually all tap their foot along.

When an event in the external rhythm occurs precisely at phase (0) of the internal oscillator, it is perceived as being exactly on time. An event occurring "before" phase (0) has a negative phase and is perceived as being early. When "after" phase (0), it has a positive phase and is perceived as late. The coupling between the two oscillations is critical for the continuous perception of rhythm, as the internal rhythm can dynamically adjust its phase and period to match the external stimulus. This allows the internal oscillator to easily track even non-isochronous rhythms. Without this coupling, even slight variances in phase or period can throw the internal oscillator off the beat. However, as long as the two oscillators can remain coupled, some sense of rhythmic structure will remain (Large & Jones, 1999).

All talk so far has been of entrainment to an external stimulus. Necessarily, in such cases only the internal oscillator can adjust directly, making the external stimulus the primary driver. However, two or more internal oscillators can also become coupled to one another so that both are able to make phase adjustments. Multiple coupled oscillators, each attending at different time periods, are termed an expectancy scheme (Drake et al., 2000). People have the ability to direct their attention to one or more of these oscillators to focus upon different aspects of one rhythmic structure. Adults most often will
synchronize with the referent level of an external stimulus, but are capable of moving to both higher and lower levels.

2.2.1 Oscillators, metrical phonology, and unit-timing

Let us look at an example of a metrical grid again. Example (10) is repeated below as (15).

\[
\begin{array}{c|c|c|c|c}
\text{Pros. Word:} & | & x & | \\
\text{Foot:} & |x| & |x| & | \\
\text{Syllable:} & |x| & x & |x| & x & | \\
\end{array}
\]

As mentioned previously, Liberman and Prince (1977) define such grids as "hierarchies of intersecting periodicities" (p. 313). This immediately raises prospects for a metrical phonology based around oscillators. In such a framework, each layer in the grid of (15) would be an oscillator. Three oscillators, one at the syllable-level, one at the foot-level, and one at the prosodic word level, would be dynamically coupled. There are several benefits to such a conception.

(1) It provides an explicit connection between the grids of metrical phonology and predictions of speech timing. The time between stressed syllables would be a function of the foot-level oscillator’s period. (2) Oscillators provide a mechanism for explaining how speech can be made to coordinate with stimuli such as a metronome when the task demands it (Cummins & Port, 1998; Port, 2003). The internal oscillators that control speech timing become entrained with the external metronome. (3) Simultaneously we can see why isochrony may be possible in speech, but only sporadically occur. Namely, each oscillator is coupled with several others in an expectancy scheme. The foot-level oscillator may have a period, but it is coupled with the syllable oscillator, which will contain syllables of varying complexity. For a series of simple CV syllables, it would be
easy to plan all required articulatory gestures within the period of the foot-level oscillator, but extremely difficult to do so when the syllable contains a long series of consonant gestures, such as the syllable *sixths*.

(4) An oscillator account of prosodic phonology also connects linguistic theory to broader psychological and neurological research. Oscillations are a well documented aspect of neural dynamics (Haken, 2002; Large, 2008), raising the hope that we might find a neural mechanism underlying this aspect of speech processing, though, at present, there is much debate about whether oscillatory timing is due to a central clock, a small number of dedicated clocks, or simply a general feature of neural computation (Irvy & Schlerf, 2008).

(5) An oscillator-based metrical phonology provides a possible explanation for the perception of stress or syllable timing through what may be termed the Directed Temporal Coupling Hypothesis for speech rhythm. The fundamental notion of directed temporal coupling is that stress-timing is perceived when primary attention is directed at the foot-entrained oscillator, and syllable-timing is perceived when primary attention is directed at the syllable-entrained oscillator. It is not necessary that either level be actually physically isochronous. Unit-timing, under this notion, is a matter of where attention is allocated.

An oscillator- or time-indexed metrical phonology also has the potential to simplify some aspects of metrical phonology. To explain, we first need to mention Cummins & Port (1998) who performed a speech cycling experiment (mentioned briefly in section 2.1.2). They found that the participants’ speech reliably collapsed to either a binary or ternary subdivision of the metronome’s beat. In the terms of the oscillator model, speakers aligned their articulatory gestures with phase (0) and phase (1/2) in the
binary case or phase (0), phase (1/3), and phase (2/3) in the ternary case. While there may be a bias towards a binary metrical perception (discussed further in Section 3.1.2), a ternary perception is certainly possible. In examples (12) and (13) above, we posited extrametrical syllables to account for words like *serendipity* so as to keep the typology of metrical feet simple. However, we can simply allow three cycles of the syllable oscillator for one foot-level oscillation to explain why the final syllable in *serendipity* is not stressed, even though a purely binary inventory of feet would require it. This “solves the problem” of words like *serendipity* without adding either new types of feet or extrametrical syllables. We know from music composed of three beats, such as a waltz, that triple meters are natural in human behavior.

While it is beyond the scope of this dissertation, an Optimality Theoretic approach to metrical phonology (Kager, 1999; Prince & Smolensky, 2004) could be an appropriate way to add time-indexation through oscillators to a phonological grammar. The sketch just provided of metrical grids as oscillators largely revolves around aligning speech to phase angles in an expectancy scheme. Work in Optimality Theory has already developed a rich set of tools around so-called generalized alignment. However, we will not tackle this here. For now, we will turn to evidence for oscillators in linguistic processing and then finally apply this work to word segmentation, which motivates the rest of the dissertation.

### 2.3. Oscillators in Linguistic Processing

Kello (2003) points out that any incremental language production system requires continual temporal coordination between a host of cognitive resources so that world knowledge, semantic content, syntactic structures, lexical items, and phonological
instructions arrive just as needed to guide complex motor systems at the stunning speed at which people are able to speak. Similarly, we must keep the items of speech input in the correct order to build the proper linguistic structures and revise them when needed.

2.3.1. Oscillators in speech production

One area in which oscillators have been incorporated into models of linguistic processing is speech production. One earlier model is OSCAR, Oscillator-based Associative Recall, developed by Vousden, Brown, & Harley (2000). This computational model takes a group of phonemes that need to be produced and attempts to put them out in the right order. It accomplishes this by having a set of oscillators act as the “context signal”. The rough idea is that the set of oscillators provides a time bed in which phonemes are placed. The validity of the model is based in large part upon its making the right type of speech errors. People not only confuse adjacent phonemes when an error is made, they confuse phonemes in similar syllabic positions, such as onsets. To use the ubiquitous example, *the dear old queen* may come out as *the queer old dean*. By more strongly activating all similar metrical positions, based upon the oscillators’ context, and introducing noise into the system, the OSCAR model makes many of the same sorts of errors that humans do.

Other oscillator-based speech production models explicitly employ foot-level and syllable-level oscillators, as opposed to OSCAR’s 32 oscillators. O’Dell and Niemenen (1999) explored the effects of such oscillators upon syllable shortening within polysyllabic words. When we compare the words *stead, steady, and steadily*, the *stead* portion in the polysyllabic words is shorter than when alone as a monosyllabic word. We already observed this phenomenon from the processing side in the *ham/hamster*
experiments by Salverda et al. (2007). O’Dell and Niemenen (1999) found that increasing the strength of the foot-level oscillator increased polysyllabic shortening, while increasing the strength of the syllable-level oscillator reduced polysyllabic shortening. Barbosa (2007) developed an oscillator model for Portuguese and was able to simulate the syllable- or stress-based rhythms of different dialects by increasing the strength of syllable- and foot-level oscillators, respectively.

Figure 1. Tilsen’s (Submitted) model of oscillator-based gestural control. (Reproduced from Tilsen (Submitted, Figure 1) by permission.)

Two very detailed production models have advanced this work all the way from the phrase level to control of articulatory gestures in speech production (Saltzman, Nam, Krivokapic, & Goldstein, 2008; Tilsen, 2009, Submitted). Figure 1 (reproduced from Tilsen, submitted) provides a succinct image of such models. In Figure 1, the oscillators are represented as sinusoidal waves. A unique wave corresponds to different levels of the
prosodic hierarchy from phrase to foot to syllable to segment to articulatory gesture. By coupling each oscillator and then placing content into this framework, both the timing and serial order requirements of speech production can be managed.

2.3.2. Oscillators in speech perception

Not as much concentrated effort has yet been spent upon explicit models of speech perception within an oscillator framework. We have already discussed the speech cycling work of Cummins and Port (1998). In that series of experiments, participants entrained their speech with a metronome and aligned their speech with the stable locations in phase space, namely phase (0), phase (1/2), phase (1/3), and phase (2/3). In music theory terms, these would be duple and triple meter.

Neuro-imaging work indicates that adults can project attention to important moments in time (Coull, 2004). This was confirmed behaviorally in a speech experiment by Quené and Port (2005). Quené and Port asked listeners to engage in a phoneme-monitoring task in which a word containing the target phoneme was placed within a list of words. Four conditions were crossed in a 2x2 designs. The target could 1) either match or not match the metrical patterns (trochees or iambs) of previous words in the list, or 2) occur at an isochronous or non-isochronous onset interval with previous words. The isochronous interval was predicted to allow participants to accurately focus their attention on an expected moment in time, resulting in faster phoneme detection. Participants were able to locate the target phoneme more quickly when it occurred at this expected moment in time, i.e., with the isochronous onset interval. Lists of words with predictable patterns of metrical structure had no effect upon reaction times. There was no significant difference if a target trochee followed a series of trochees or a series of iambs. This
matches with the attentional aspects of Large and Jones’ (1999) model of time perception. The earlier items in the isochronous condition were able to heighten attention at a predicted moment in time. When the target phoneme occurred at that point, they noticed it more quickly than otherwise.

Other intriguing support comes from Astheimer and Sanders (2009). They were not studying rhythm, but instead asking if listeners selectively attended to word onsets. They asked participants to listen to an audio recording of a novel by its author in which they had inserted attentional probes either at word onsets or in control conditions. As measured through event related potentials, listeners showed greater attention specifically at word onsets. As they point out, their study does not provide an explanation of how attention becomes focused upon word onsets. Per the Large and Jones model, rhythm could give us a way to manipulate the temporal location of attention.

If attention can be placed upon word onsets using a temporal pattern, this could be a possible word segmentation cue. As mentioned at the end of Chapter 1, the Saffran et al. (1996) artificial language uses an entirely isochronous rhythm, so, if the rhythm can be aligned with the word onsets, then learning could occur without any statistical calculations. In Chapter 3, we will put all of this work we have done on statistical learning, word segmentation, and time perception together and present the experimental research to be done.
CHAPTER 3: ASSESSING TIME AND STATISTICAL LEARNING IN WORD SEGMENTATION

3.1. Finding Words Through Temporal Expectation

3.1.1 Review of statistical learning, word segmentation, and attention in time.

Chapter 1 introduced the potential importance of statistical learning to linguistic processing. Researchers have presented evidence for statistical learning at phonetic, phonological, lexical, syntactic, and semantic levels. Corpus studies additionally indicated that robust statistical patterns exist in many areas of language, so that, if we have the abilities to find these patterns in natural interaction, much linguistic structure can be discovered through statistical calculation. Much of this work has explored the possible usefulness of statistical learning in the problem of word segmentation, or finding words in a continuous speech stream. While a series of experiments has supported the notion that statistical calculations can be used to find words, when a competing linguistic cue, such as stress, lengthening, or phonotactics, is added to an artificial language, infants and adults more often than not appear to use the other cue and not statistics.

This has often been argued to show that greater weight is attached to certain cues. However, many such cues are available immediately, while a statistical pattern can only unfold gradually over time. Thus, the use of one cue over another could be due to temporal priority as much as strength.\(^{11}\) More broadly speaking, we might characterize the behavior as searching for structure in the speech stream immediately and continuously.

\(^{11}\) Temporal priority does not rule out the notion of greater and lesser weight, of course. Both could be in play.
In the latter sections of Chapter 1, we also explored the possibility that a statistical boundary could mark a prosodic boundary either in addition to or instead of a word boundary. This was partly motivated by the wide evidence of the relationship between prosodic boundaries and word boundaries and partly by the suggestion that statistical unpredictability, which occurs at the end of each word in a Saffran et al. type artificial language, could trigger a subjective perception of increased duration. Increased duration, in the form of lengthening at prosodic boundaries, is in turn a documented cue for word segmentation.

Due to the notion that there might be a relationship between statistical word segmentation and prosodic boundaries, Chapter 2 turned to a more detailed look at prosody in language. Most of the rhythmic structures in stress languages around the world can be compactly described using the terms of metrical phonology, such as binarymetrical feet on a grid of strong and weak pulses. English has a dominant structure of trochaic strong-weak feet, though other feet are possible, such as a monosyllabic word. A single syllable will be more acceptable the more it approximates the duration of a full binary foot.

Finally, the latter part of Chapter 2 introduced the notion of time perception as a set of oscillators with a period and phase. The more predictable and structured a temporal pattern is, the more attention will focus on particular points in the phase space. We also saw behavioral evidence that attention can be directed towards an expected moment in time. This opens the possibility of cueing the location of word onsets through manipulation of temporal structure.
3.1.2. The impact of isochrony on listening to speech

Segmenting speech, particularly for the sort of artificial languages that we are focused on in this dissertation, is essentially a grouping exercise. The listener is attempting to find the right way to group syllables so as to form words. The challenge, of course, is that the problem is unconstrained. One might group any number of syllables together and those groups might start at any point. A random grouping would appear hopeless. This is where rhythm can help – by constraining the possible structures to consider.

When people hear a repetitive time sequence, they naturally impose a rhythmic structure with 2 or 3 units per grouping. The physical *tick-tick* of a clock is heard as *tick-tock*. This is a robust finding in music theory and psychology (Bergeson & Trehub, 2006; Lerdahl & Jackendoff, 1983\(^{12}\); Phillips-Silver & Trainor, 2007; Repp, 2007) and has been known for a century (Bolton, 1894, as cited in Large, 2008). To remove prosodic cues, such as stress and final vowel lengthening, Saffran et al. (1996) created an entirely isochronous language. However, a completely isochronous pattern will draw rhythmic attention by strongly supporting a rhythmic oscillator (Drake et al., 2000). Adjusting the phase of one oscillator to remain synchronized with another may be automatic in the sense that adjustments happen both immediately and seemingly beyond conscious control (Repp, 2005). If the isochronous nature of the artificial language attracts rhythmic attention, as isochronous sequences typically do, the listener is highly prone to impose either a binary or ternary grouping. This immediately constrains the number of models.

---

\(^{12}\) This is Lerdahl and Jackendoff’s Metrical Well-Formedness Rule 3.
being considered. The only parameters to discover now are the alignment points between syllable and metrical structure and the choice between groups of two or groups of three.

We can interpret this problem with the concept of an oscillator- or time-indexed metrical phonology as suggested earlier in Section 2.2.1. To segment the entirely trisyllabic language completely, the goal is to entrain a foot-level oscillator with word onsets. The foot-level oscillator is in turn coupled with a syllabic oscillator that will cycle three times for each period of the foot oscillator (Figure 2). As in Figure 1, the oscillator is represented as a wave. The foot-level oscillator is represented as cycling 3 times in the diagram, while the syllable-level oscillator cycles 9 times, 3 times within each foot.

<table>
<thead>
<tr>
<th>Foot</th>
<th>Syllable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. A syllable-level oscillation repeats three times for each foot-level oscillation.

While listening to the isochronous artificial language, the participant will become entrained with the speech. From an oscillator standpoint, this entails that the phase and period of the participant’s internal oscillator be in coordination with the phase and period of the external speech stimulus. The participant should entrain easily with the syllable-level oscillation within the stimulus, since the participant can simply move from adjacent syllable to adjacent syllable. The foot-level oscillator has the difficult problem, since it must find the correct coupling between both the external stimulus and the internal syllabic oscillation.\(^\text{13}\)

\(^\text{13}\) The syllabic oscillator must also couple with both an external stimulus and the internal foot oscillation. The problem is simplified for it, however, since the external stimulus at the syllable level simple moves from syllable to syllable. As each syllable has a single vowel, the coupling only needs move from one increase in acoustic energy (at the vowel) to the next increase in acoustic energy.
In our review of metrical phonology, we saw that a content word aligns with at least a foot, so the foot-level oscillation is key to word segmentation. If phase(0) for the participant’s foot-level oscillation matches the word onset of each isochronous word, then this might cue a prosodic boundary. Moreover, if participants impose a metrical structure upon the speech based upon the isochronous rhythm, which amounts to imposing a certain coupling with the internal syllable oscillation, then this grouping could assist in clustering syllables of a word together.

Of course, a problem with such an account is that there are multiple ways to impose a metrical grid upon the artificial language. One such way – groups of three beginning at a word onset – segments the language entirely. However, if the alignment mapping is off, such as starting on the second or third syllable of a word, or the grouping is incorrect, such as employing groups of 2 not 3, then the words will not be successfully found. Is there a way to overcome these problems?

3.1.3. How to segment using the cue of timing

To build a proposal, I will focus on the stimuli from Saffran et al. (1996). The essential idea is that the participant will try to maximize the amount of structure found in a parse. In this sketch, two rough metrics can be used: 1) Groupings that are repeated are maintained as a viable parse. 2) The participant wishes to minimize the amount of input that is not repeated. Metric 1 is straightforward. If the parse yields groups that continually recur (i.e., the groups make good predictions for the future input), then structure is being found. Metric 2 is present to reduce the amount of input that is unexplained.

A metric of repetition has been employed for several reasons: 1) Previous research has indicated that some speech streams can only succeed with repetition of
critical items (Bonatti et al., 2005). Endress et al. (2009) suggest one learning mechanism could be specialized to look for repetition in speech. 2) Perruchet and Vintner (1998) developed a computational model, PARSER, that works by dividing a speech stream into chunks, and then keeping those that repeat in memory, while letting others decay. This model successfully finds the words in the Saffran et al. (1996) language. 3) Repetition is a simple way to conceive of the participant predicting future speech. The participant predicts simply that they will hear the same thing they just heard. When that predicted input does repeat, the prediction is confirmed.

In the following, we will use a simple ratio to measure “finding structure”.

\[
\text{Parsing Success (PS)} = \frac{\text{# syllables repeated within a grouping}}{\text{total # syllables}}
\]

However, this formula is simply to give an impression of the repetition metrics put to use. Groupings that are repeated will increase PS, but if the repetitions are only a small part of the total input, then the parse will be penalized. There is no claim that this formula is the ideal objective function for model building.

As discussed in Chapter 1, Saffran et al. created six tri-syllabic words: babupu, bupada, dutaba, patubi, pidabu, and tutibu, which were randomly concatenated except that no word immediately repeats itself. Example (16) represents such a stream orthographically with syllable numbers added below the text for reference:

(16) bupadatutibabupupatubibupadabupupidabututibu…

Segmenting words in a case such as this essentially means forming groups of three adjacent syllables. If one has a way to pick the right onset, the next word in such a language will always be an exact moment in time away, namely the duration of three
syllables. This is marked in (17) with **bold** text representing each word onset. This is actually the perfect parse and finds the causal structure in the model, i.e., the complete list of words the speaker is using to produce the speech. However, the participant does not know this until substantial time passes. Example (18) lists the words found in the speech encountered so far.

(17)  
\textbf{bupadatibubabupupatubibupadababupupidabututibu…}  
\textbf{1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \ 10 \ 11 \ 12 \ 13 \ 14 \ 15 \ 16 \ 17 \ 18 \ 19 \ 20 \ 21 \ 22 \ 23 \ 24}  

(18)  
bupada, tutibu, babupa, patubi, bupada, babupa, tutibu

In (17) three words are already repeated (\textit{bupada} starting at syllables 1 and 14, \textit{tutibu} starting at 4 and 22, and \textit{babupa} starting at 7 and 16), which would both reinforce memory traces for those words and indicate that structure is being identified. To apply the PS metric, 3 trisyllabic words were repeated for a total of 18 syllables grouped and then repeated in the same groups later. 24 syllables have been encountered so far, so \(PS = \frac{18}{24} = 0.75\).\(^{14}\)

If the listener imposes groups of three, but starts at an incorrect alignment point, such as the second or last syllable, no robust pattern emerges. Example (19) displays the imposition of tri-syllabic groupings starting in the "wrong place," in this case a medial syllable, while (20) lists the possible words that emerge from such an interpretation.

(19)  
\textbf{bupadatibubabupupatubibupadababupupidabututibu…}  
\textbf{1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \ 10 \ 11 \ 12 \ 13 \ 14 \ 15 \ 16 \ 17 \ 18 \ 19 \ 20 \ 21 \ 22 \ 23 \ 24}  

(20)  
padatu, tibuba, bupupa, tubupu, badaba, bupupi, dabutu

\(^{14}\) One can debate the precise way to count syllables and groupings, such as whether the first grouping should count, etc. Truly, however, the PS measure is only for illustrative purposes. Any plausible metric for one that could actually be used by participants would be far more dynamic and independently justified.
There is nothing inherently wrong with the words of (20). However, no repetition has occurred at all yet. PS for the parse of (19) and (20) is 0. No pattern has yet been seen with interpretation (19). It will grow from 0 as more data is acquired, depending on the combinatorial dimensions of the speech.

Finding structure in the language requires knowing the correct number of syllables in a group as well as the correct alignment point. In (21), disyllabic groupings are imposed, instead of the correct trisyllabic groupings, and the resulting words are displayed in (22).


(22) bupa, datu, tibu, babu, pupa, tubi, bupa, daba, bupu, pida, butu, tibu

This set of words has a bit more structure than (20) in the sense that we have two repetitions with *bupa* and *tibu*. An interpretation with two repetitions does not appear substantially different than the three repetitions found in the correct analysis (17). However, the disyllabic analysis leaves 16 of 24 syllables without any reinforcement. We can calculate PS here as 8 syllables repeated over 24 for a value of 0.33. If a speaker's goal is to find structure in the stimulus (Gomez, 2006), much more structure is to be found in the trisyllabic analysis of (17).

What we need at this point then is a method of converging on the interpretation with the most structure. First, recall that using the rhythm of the speech has greatly constrained the model choices. There are essentially only 6 possibilities: disyllabic groups starting at three different alignment points or trisyllabic groups starting at three alignment points. We must suppose that the participant works through these 6 possibilities and finally concentrates on the one that maximizes PS. Section 1.3.3
reviewed previous research that indicated people can consider multiple models simultaneously, though the limits of this are largely unknown. At the same time, also from Section 1.3.3, participants might consider some models before others. Gomez (2006), for instance, argued that adjacent statistics are considered before non-adjacent ones. In the case of word segmentation, an adult participant may be strongly biased towards disyllabic feet, since this is the primary metrical foot of English.

We have largely discussed groupings in this proposal, so it is important to remind ourselves of the importance of time, not simply grouping. The time element helps in two ways. First, it is the isochronous nature of the language that draws rhythmic attention. Secondly, rhythmic attention will be drawn to two periodicities in the language. A speech stream sample, (23), with the correct interpretation is copied again for discussion.

(23)  \[\text{bupadatutibubupupatubibupadababupupidabututibu…}\]

The first periodicity here occurs from one syllable to the next. Each syllable’s vowel onset occurs at a highly predictable time with which an attending oscillator should easily become entrained. A second higher-level periodicity occurs with each word onset. As the interpretation begins to segment words, it will discover word onsets uniformly occurring at a moment in time exactly predicted by the previous interval between word onsets (the interval between word onset \(W_i\) and \(W_{i+1}\) is identical to the interval between word onset \(W_{i-1}\) and \(W_i\)). This should set up a temporal expectancy for each word. Indeed, once this temporal expectancy is learned, nothing else is needed to find word boundaries. The listener only needs to stay "on the beat". This beat is based upon the isochronous nature

\[\text{15 Differences in spaces in the diagram are present to align with the font above. Temporal duration is intended to be uniform between syllables.}\]
of word onsets in the stimulus, but that is not the nature of the expectancy. The expectancy is simply that a word onset will occur close to the phase (0) marker of the attending rhythm that follows the word-level (or foot-level) periodicity, which, in more natural speech, could change dynamically.

### 3.2. Is Isochrony a Confound in Statistical Learning Experiments?

The goal of the previous section was to build a plausible case that segmenting an artificial language might be accomplished through applying metrical structure to the speech stream. More precisely, through a search of a small number of models, the participant could become entrained with the isochronous word onsets of the language. Doing so would then mark either morphological word onsets or prosodic onsets that are then mapped to word onsets. If this is possible, it opens up the possibility that the isochronous rhythm of the language in many word segmentation experiments is a confound for providing evidence of statistical learning. The participants may be learning the language through temporal expectation, i.e., that critical events in linguistic structure will be aligned with focal points in time. Perhaps Saffran et al. (1996) does not provide evidence for statistical learning at all?

We have two additional pieces of evidence that this might be the case. Dilley and McAuley (2009) conducted an experiment with an ambiguous series of words that could be grouped in at least three ways. For instance, the series *note* - *book* - *worm* could be heard as three monosyllabic words, or as *notebook* and *worm*, or as *note* and *bookworm*. They found that such a series can be disambiguated by temporal patterns that occur “distally” (in this case a few syllables before) the ambiguous point (Dilley & MacAuley, 2009). A rhythmic pattern will be carried through a speech stream so as to divide later
portions of the stream. Dilley and MacAuley conclude that “the present work suggests that distal prosodic information several syllables in advance of proximal segmental and coarticulatory cues can potentially influence lexical processing, so that prosodic information may be more useful and/or influential in processing than previously believed” (p. 307).

A second piece of evidence arises from a condition in which statistical learning is not pitted against another cue, and so should not be “outweighed” for segmentation, but the participants nevertheless fail to learn the language. Tyler and Johnson (2006) created an artificial language that included both disyllabic and trisyllabic words, but maintained clear dips in transitional probabilities between words. The forward transitional probability between each syllable within a word was 1.0, while between words it dipped to 0.33. The infants in the experiment, however, did not display any preference for either words or part-words in testing, suggesting they could not discriminate the two.

In the Tyler and Johnson research with infants, all syllables had a uniform duration, as is customary to remove stress and lengthening cues. This implies that words in a mixed length language are not merely of different length in syllables but also different in time. If each syllable were to be 300 ms, then disyllabic words would be 600 ms, while trisyllabic ones would be 900 ms. While we do not know the exact parameters of entrainment, this certainly would make a rhythmic solution far more difficult than in a language with words that are all of the same duration. The previous proposal for finding structure through temporal grouping only worked because it did not matter if one started on a first syllable in word 1 or word 15. The structure is the same and one could simply look for a relationship at a certain temporal location in the future. But with the word
durations varying, a temporal guess of either 600 or 900 ms (2 or 3 syllables) will parse the language differently based upon where one starts in the stream.

3.3 Research Questions, Hypotheses, and Design

Tyler & Johnson (2006) provide a case in which statistical cues are similar to those in other word segmentation experiments, but learning does not occur. Simultaneously, the rhythmic cue available in the previous experiments where participants did successfully segment is not available and learning does not occur. Let us state, therefore, the following very bold Time Not Statistics hypothesis:

**Time Not Statistics**: All word segmentation in a language with isochronous syllables is accomplished through Expected Temporal Alignment (ETA), because there is an expectation for certain events to occur at predictable moments based on the temporal structure of the speech. No learning occurs due to statistical calculations.

The reason the Time Not Statistics proposal is very bold is that there have been many other apparent demonstrations of statistical learning that have no clear connection to temporal patterns. Section 1.3.1 covered claims for statistical learning of items such as word-object labels and hierarchical phrase structure. It is not clear how ETA would account for such results. A more tempered hypothesis would therefore propose that both are possible cues, the Time And Statistics hypothesis.

**Time And Statistics**: Participants can use statistical uncertainty, such as a decrease in transitional probability between syllables, and ETA to segment words in speech.

To test these two hypotheses, we need a way to remove and restore one of the cues. Tyler & Johnson’s (2006) methodology offers one possible way to do this. Their artificial language removed a consistent rhythm between words, and hence the participant’s ability to form useful temporal expectations for word onsets, by changing
the number of syllables in a word, but leaving the durations of all syllables the same. This resulted in words of different durations. If we were to modify the durations of the syllables such that trisyllabic words were of the same duration as disyllabic words, then all words would be of the same duration. This might then allow a foot-level oscillator to entrain with word onsets and restore learning. The next section works through the overall logic of the experiments with detailed methods and results in the following Chapters 4-7.

3.3.1. Outline of the experiments with predictions

A series of experiments investigates the relevance of temporal expectancy for word segmentation, using the results of Tyler and Johnson (2006) as the organizing principle. As discussed, Tyler and Johnson discovered that infants have difficulties segmenting speech when a language is composed of both di- and tri-syllabic words. This result is not predicted by a simple statistical learning account, since the transitional probabilities dip in such a language just as they do in a language such as that of Saffran et al. (1996). However, by changing the number of syllables in words while keeping all individual syllable durations the same, Tyler and Johnson changed the timing of word onsets as well. Experiments 1 and 2 attempt to duplicate these findings with adults. Experiment 3 tests the listener's timing-based segmentation abilities by manipulating the durations of syllables to restore the temporal regularity of word onsets while preserving the word variability of Tyler and Johnson’s language. Table 2 provides a snapshot of all three experiments.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Brief Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>All disyllabic language with isochronous syllables and words</td>
</tr>
<tr>
<td>2</td>
<td>Mixed di- and tri-syllabic language with isochronous syllables</td>
</tr>
<tr>
<td>3</td>
<td>Mixed di- and tri-syllabic language with isochronous word onsets and consistent intra-word syllable durations.</td>
</tr>
</tbody>
</table>
**Experiment 1** extends the Saffran et al. (1996) results to a disyllabic language, as previous research employed trisyllabic words. Both the ETA proposal and statistical learning support segmentation in this case and so the learning level should be quite high. This experiment serves as a comparison point for the following more interesting experiments. **Experiment 2** largely replicates Tyler & Johnson, but with adult participants. In this experiment, syllabic rhythm is highly structured with isochronous syllable onsets, but the location in time of word onsets fluctuates due to variability in word length. If only ETA is a cue to segmentation, learning should not occur, but, if only statistical learning is a cue, learning *should* occur. If both temporal expectation *and* transitional probabilities are available cues for segmentation, the predictions are more complicated.

As discussed in Chapter 1, participants will begin to segment the speech immediately so that they may be drawn to using the temporal structure before the statistical pattern has unfolded. Following Gomez’ (2006) dynamically guided learning, as well as the garden path concept from Gebhart et al. (2009), they may remain with this cue throughout the training period unless insufficient structure is found (see Section 1.3.1 for discussion). In Experiment 2, the only reliable temporal structure is from syllable to syllable and not from word to word; therefore, they should be more prone to abandoning any attempted parses using the rhythmic cues and move to statistical cues. In experiment 3, to be discussed below, temporal structure is richer and so they may never move to statistical cues. Recall that these predictions are all under the assumption that both time and statistics are valid segmentation cues.
Of course, Tyler and Johnson (2006) found that infants were not able to distinguish words and part-words in a language such as Experiment 2. If adults do learn, infants must lack some knowledge or ability that adult speakers have. One possible reason for the infant’s failure is because of the lack of a word-level rhythm. One can examine the importance of word onset timing by regularizing word onsets again; i.e., by using a language with both disyllabic and trisyllabic words, but where the length of vowels in the disyllabic words is modified so that the entire word takes the same amount of time as the trisyllabic one. This is our Experiment 3. Let us break down what learning in Experiment 3 would mean for the Time Not Statistics and Time And Statistics hypotheses.

The predictions are straightforward for the Time Not Statistics hypothesis. Learning should not occur in Experiment 2, since there is no word-to-word temporal cue, but it should occur for Experiment 3 where the word-to-word temporal cue is restored. For the Time and Statistics hypothesis, the temporal cue will again be available before the statistical structure. Unlike Experiment 2, however, in Experiment 3 the temporal structure entirely segments the speech with no stray syllables. Even if statistical cues can be used to segment speech in real-time, the learning from temporal cues would hide that. However, since statistical cues are available in Experiment 2, we should see some learning there. The only reason learning would not occur in Experiment 2 if statistics are a valid cue is if some other cue is available prior to the statistical one and, moreover, appears to find a robust structure that happens to be incorrect.

If learning is restored in Experiment 3 with its isochronous word intervals, it suggests that simple timing of word onsets can be a cue to word segmentation. If both
statistical learning and ETA are cues, it suggests a high level of learning again as in Experiment 1; however, ceiling effects could obscure any such inferences.

Two more matters should be discussed before moving to the experiments. The first topic is what it means to “learn” the words of an artificial language. The basic criterion for establishing learning has traditionally been a statistically significant selection of words over foils. This is a valid measure, but of course many different behaviors could yield such a result. To see which behavior triggered the selections, it is necessary to look beyond the group mean for accuracy and examine the exact words selected to see if they share any particular properties. If any solution found the entire words, such that participants perform at ceiling, then the precise selections should be largely uninformative, since the participants would select all words evenly.

However, if either transitional probability or the temporal expectation for word onsets is a prosodic cue rather than a complete lexical grouping, there could be interesting patterns in the word selections. Chapter 1 raised the possibility that the increase in uncertainty at the end of a word will be subjectively experienced as an increase in duration. Lengthening is a characteristic in speech right before a prosodic boundary, i.e., lengthening signals an end. Assuming the participant applies their native language’s phonology to the speech stream, this would most naturally be construed as the end of a binary foot, which focuses the participant on only the latter two syllables of a trisyllabic word in the artificial language, not the whole word. This could then be signaled in the testing phase if participants erroneously choose foils that include that pair of syllables, but not foils that do not include that pair.
Finally, one “loose cannon” in predicting experimental results is how participants in Experiment 3 will interpret the junctures of trisyllabic words and disyllabic words. Unlike Experiments 1 and 2 (and unlike Saffran et al. or Tyler and Johnson), Experiment 3 contains syllables of different durations so as to restore the isochronous word intervals. When trisyllabic words follow trisyllabic words, the syllables will be of identical durations, and the same is true when disyllabic words follow disyllabic. However, if a disyllabic word follows a trisyllabic, the first syllable of the disyllabic word will be longer than the last syllable of a trisyllabic word. The result would be something like (24).

(24) kadudadi:gu:

It is possible the di: syllable would be interpreted as lengthening or the strong syllable of an iambic foot, resulting in a perceived foot dadi:. Similarly, when a disyllabic word is followed by a trisyllabic as in (25), the last syllable of disyllabic word will be longer than the first syllable of the trisyllabic.

(25) di:gu:kaduda

This could be interpreted as stress in a trochaic foot, resulting in a perceived foot gu:ka. We can look for evidence of these complicating factors in the pattern of selections.
CHAPTER 4: EXPERIMENT 1 – LEARNING AN ALL-DISYLLABIC LANGUAGE

4.1. Methodology

Experiment 1 examines whether adults can find words in a speech stream when all words are disyllabic. This experiment extends the results of Saffran et al. (1996) who employed a trisyllabic language with adults and Tyler & Johnson (2006) who employed an all-disyllabic language with infants. In both previous research studies, the participants were able to find words and it is expected that adults will succeed as well with an all-disyllabic language. This experiment serves mostly as a point of comparison for Experiments 2 and 3.

Word learning will be tested by having adults first listen to an artificial language in a familiarization stage and then take a forced choice test that is composed of pairs of words from the language and foils. The foils are composed from syllables that occur together during training but across a word boundary. These boundary words then occurred during familiarization, but not with the same statistical pattern as a true word. Henceforth, we will use the term “l-word” to indicate a word in the artificial language, “boundary word” to indicate a word that is composed of syllables that occur across an l-word boundary, and “word” to indicate any type of word. Two languages will be employed in this experiment as a control (Gomez, 2002; Gerken, 2004). Language A is composed of 6 l-words, tested against 6 boundary words. Language B takes the boundary words from Language A and uses them as the words of the language. If the words of the two languages do not have accidental properties that trigger their selection, results should be comparable in both languages.
4.1.1. Participants.

Twenty-four students at the University of Hawai‘i participated in the experiment for either course credit or $5. Twelve were assigned to language A and twelve were assigned to language B. All participants self-reported normal hearing and that English was one of their native languages. The entire study occurred with approval from the university's Committee for the Protection of Human Subjects.

4.1.2. Stimuli.

While the goal for Experiments 2 and 3 is to extend the results of Tyler and Johnson (2006), there is one drawback to using their precise languages. They employed consonants with a variety of manners of articulation. However, fricatives are inherently longer than stops in duration and we wish to control the timing of vowel onsets very precisely. If we reduce the fricative to be the same duration as a stop, it no longer sounds like a fricative.

Therefore, it was decided to follow the model of the Saffran et al. experiments and employ an artificial language containing 6 disyllabic l-words with only stops as consonants. Each word was composed of two CV syllables containing one stop and one vowel. The phonemes used were: [p], [t], [k], [b], [d], [g], [i], [a], and [u]. The six l-words were randomly concatenated together with the exception that no l-word could repeat itself. The transitional probability between syllable 1 and syllable 2 in an l-word was exactly 1.0, since a 2nd syllable always occurs after its 1st syllable mate. Any l-word can be followed by any of the other five l-words (a word cannot follow itself), and so the transitional probability between syllable 2 of one l-word and syllable 1 of the next l-word is 0.2. The six l-words in Language A were: *digu, kuba, bupi, kadu, gaki,* and *tipa.*
Voiceless stops occur as the initial segment in three words and voiced stops as the initial segment in the other three. Careful attention was paid to this since voiceless stops are potentially a word cue in English. In many dialects of English, including the American English of the participants, a voiceless stop phoneme is aspirated when it is syllable-initial at the beginning of a metrical foot (see Section 2.1.2 for a discussion of metrical phonology), but not aspirated when in an unstressed or non-initial location. The /t/ phoneme in particular has several allophones in Mainstream American English. Compare the words *atom* and *atomic*.

(26) atom

[æ.ɹʌm]

|x |  
|x. x |

(27) atomic

[ə.tʰʌ.ˈmɪc]

|x |  
|x | x  
|x | x x |

The key observation is that, when /t/ begins the metrical foot as in *atomic*, it is aspirated but takes several non-aspirated allophones in other environments. This then suggests that one could posit that a foot begins during the artificial language stream whenever an aspirated voiceless stop occurs. One possibility would be to not use voiceless stops in the artificial language, but that would only leave three stops for the language’s syllabary. As voiceless stops have been used commonly in previous research, the decision was to use voiced and voiceless stops, but ensure that they were distributed across l-words and across syllables so that they could not reliably signal l-word onsets. The speech synthesizer, discussed below, generated all aspirated voiceless stops.
The vowel inventory was also distributed across the l-words, so that [i], [a], and [u] each occurred in the initial syllable twice. Finally, no vowels were repeated within an l-word in case that was a particularly salient pattern.

Boundary words were created using the last syllable of a word and the first syllable of a different word. The selected boundary words were piku, badi, duga, kika, guti, and pabu. Again, there were three initial voiceless stops, and a distribution of the initial vowels. Language A's l-words and boundary words were switched for Language B. Table 3 list all l-words and boundary words for Experiment 1.

<table>
<thead>
<tr>
<th>L-Words</th>
<th>digu</th>
<th>kuba</th>
<th>bupi</th>
<th>kadu</th>
<th>gaki</th>
<th>tipa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boundary Words</td>
<td>piku</td>
<td>badi</td>
<td>duga</td>
<td>kika</td>
<td>guti</td>
<td>pabu</td>
</tr>
</tbody>
</table>

A single 7-minute block was created through random concatenation (without repetition) of each language’s words. All speech was produced using the MBROLA speech synthesizer (Dutoit, 1997) with the us1 diphones database. Every stop consonant was instructed to be 30 ms in duration and every vowel 300 ms in duration. This resulted in absolutely uniform 330 ms syllables and 660 ms disyllabic words. All speech was produced with a fundamental frequency of 175 Hz with no instructions to fluctuate intensity. The output of the synthesizer was recorded as a .wav file (16-bit; 44,100 Hz) using Audacity (Mazzoni, 2008).

For the testing stage, the individual word tokens were generated again by the MBROLA synthesizer. Each of the l-words and boundary words were paired

---

16 After all experiments were completed, it was noted that all syllables were 330 ms in duration, but the duration of the consonants could be longer than when aspiration was added.
exhaustively, creating 36 test items. If nothing is learned, participants should choose l-
words and boundary words at 50% on average, or 18 correct.

4.1.3 Procedures

Participants were told they would listen to a 'nonsense' language composed of
words but no grammar or meaning. Their task was to discover the words of the language.
After being guided through a set of instructions, they listened to the training language in a
single 7-minute block. The first and last 5 seconds of each block were faded in and out,
respectively, so that the starts and stops of training would not provide a cue to
segmentation. When the language was complete, participants hit a key to move to the
testing phase.

The test instructions were to hit a designated key on the left side of the keyboard
<\textit{a}> if the first word they heard was a word in the language, and to hit a designated key
on the right side of the keyboard <\textit{k}> if the second word was a word in the language.
There was no practice session. Each test item was composed of a word and a boundary
word, each of duration 660ms, with 1 second of silence between a word and boundary
word. After playing a word pair, the experimental software waited until one of the two
designated response buttons was pressed before moving to the next item. Items were
presented randomly for each participant. After presenting all 36 pairs of test words, the
experiment ended.

All experimental sessions occurred in a quiet room with the participant wearing a
headset and took about 10 minutes to complete an entire session. After a participant
finished, the researcher would speak informally with the participant about their
perceptions of the experiment.
4.2. Results and Discussion

Results. None of the 24 participants in Experiment 1, the 12 in language A or the 12 in language B, were excluded during analysis. As the testing phase included 36 binary forced-choice questions, a chance response would be a score of 18 or a proportion correct of 0.5. Participants in both languages achieved a mean above chance (1A Score: $M=20.92$, $SD=5.28$; 1A Proportion Correct: $M=0.5810$, $SD=0.1457$; 1B Score: $M=21.58$, $SD=4.56$; 1B Proportion Correct: $M=0.5995$, $SD=0.1267$). See Figure 3 for means and confidence intervals of both experiments.

![Exp 1A / 1B Mean plus Confidence Interval](image)

Figure 3. Mean proportion correct with confidence intervals for Experiment 1.
Two steps were taken to assess whether the results followed a normal distribution. First, the density function was plotted for visual inspection for both Experiments 1A and 1B (Figure 4).

While the resulting density functions do not follow a perfectly smooth normal curve, some variability is to be expected with only 12 data points. A Shapiro-Wilk test for normality indicated that neither 1A nor 1B significantly violated normality (1A: $W=0.9357$, $p=0.4442$; 1B: $W=0.8845$, $p=0.1002$), though 1B does have a $p$-value of 0.1, nearing significance.

The variances between 1A and 1B were also insignificant ($F(11,11)=1.3407$, $p=0.6351$). This allows us to use a one-way ANOVA to assess whether the means of 1A differed from that of 1B. The means were not significantly different ($F(1,22)=0.1095$, $p=0.7439$) between the two languages. Since the non-normality of 1B was near
significance, a non-parametric Kruskal-Wallis test was run as well. The means were not significantly different on this analysis ($\chi^2(1)=0.0842, p=0.7717$).

Since there was no significant difference between the means of 1A and 1B, they were collapsed together for further analysis. A Shapiro-Wilk test confirmed that the consolidated results followed normality ($W=0.908, p=0.9101$). The results for Experiment 1 were over the chance level of 0.5. ($M=0.5903, SD=0.1344$). A one-sample t-test demonstrated that the means were significantly different from chance ($t(23)=3.2905, p=0.0032$). This result is comparable to the results of Saffran et al. (1996). Their part-word test obtained a score of 22.3 (65%) while the current experiment obtained a score of 21.2 (59%). Among various differences between the two experiments, the current participants only went through training for 7 minutes, as opposed to 21 minutes in Saffran et al. Additionally, Saffran et al. used part-words made of 2 syllables from an l-word and some other syllable, not necessarily crossing a boundary, which means that the part-words of Saffran et al. were never heard in their entirety during training. Boundary words provide a more difficult discrimination task.

Discussion. There were several purposes for Experiment 1. First, previous work had not used an all-disyllabic language, so the results of Experiment 1 confirm that participants can segment a disyllabic artificial language. The use of two languages, A and B, also provided evidence that the results were not due to unanalyzed features of the selected words chosen, but instead some feature common between the two languages: a rhythm, statistical patterns, or something else. Both primary hypotheses under consideration, rhythm and statistical learning, predicted learning in Experiment 1, and this was confirmed. Finally, as the results were comparable to those of Saffran et al.
(1996), it suggests that the current methodology is also testing similar abilities to those tested in the Saffran et al. paper.

One intriguing question, however, is this: why is the proportion correct not higher than 59%? Saffran et al. (1996) report a slightly higher number of 65%, but that number means over a third of the test responses are incorrect on average. The statistical structure, however, identifies the words completely, not partially. There are two main types of explanation for the moderate success in the task (instead of very large success): 1) The participants segment the language completely, but fail to show that on the test due to perceptual confusion of synthesized syllables, memory load, inability to recall the segmented words, boredom during the test, etc. 2) The participants only segment the language partially, which may be compounded by the memory/attention-type constraints.

Evidence against explanation 1) comes from other segmentation experiments in which people do better. For instance, in a second experiment Saffran et al. (1996) lengthened the final syllables of three of the l-words in addition to leaving the transitional probability cues for all l-words. The proportion correct increased to 80%. If participants’ moderate success in the distribution-cue-only condition was due entirely to confusion or poor recall during the test, then that should happen to a similar degree for the participants in the lengthening condition of Saffran et al.

One could hypothesize that some participants do segment the Experiment 1 speech completely using distributional cues, such that they get a very high score, and others do not segment it using distributional cues, such that they get a score around chance. As a whole then, the group would achieve a score around 59%. If that were the case, however, we should expect a bimodal distribution – one distribution of high scorers
and one of chance scorers, composed of the segmenters and non-segmenters respectively. This does not match the results of Experiment 1. A histogram provides little evidence of a bimodal distribution of high scorers and low scorers (Figure 5).

![Histogram of Experiment 1 Responses](image)

Figure 5. Histogram of proportion correct for Experiment 1.

The more likely explanation of the moderate success in Experiment 1 is that participants have only partially segmented the language, not that they segmented it completely and fail to show that. This is somewhat odd if statistical distributions are the only available cue since the statistical pattern marks the words completely. It would be logical, however, if there is a competing noisy cue.

Recall that there are isochronous intervals between syllables and word onsets. If the participants are devoting rhythmic attention to the isochronous syllables and then imposing two- or three-unit groupings, they will divide up the speech with differing success, depending on the groups they try. Those syllable-based rhythmic groups could
go across l-word boundaries or remain within l-words. Such mappings are available before the statistical structure has unfolded. The participants who by chance imposed binary groupings upon the true l-words would find them later reinforced by the distributional cues and therefore score highly. Those who by chance imposed binary groupings on boundary words may have a more difficult time backing off their incorrect segmentation (Gebhart et al., 2009) and only display moderate success.

Of course, if participants can find the larger isochronous interval between word onsets, then this should help find words in the language. Section 3.1.3 proposed a possible way in which participants might search through the possible models of rhythmic structure and find the one with the maximum structure. The results of Experiment 1 (as well as those of Saffran et al., 1996) suggest that participants cannot reliably do this. If they could, we would expect higher scores, such as were seen in the lengthening condition of Saffran et al. In summary, one possible reason that participants do not achieve very high scores is that they do group the speech based upon isochronous syllabic intervals, but do not easily find the isochronous word intervals.

One alternative account for the moderate learning of Experiment 1 could depend upon some characteristic of the participants. For instance, perhaps there is a distribution to the amount of attention the participants devote to the speech during the training period. We have no assessment of their attention levels during the 7 minutes of training. There would be a distribution of attention to the task just as much with the Saffran et al. lengthening condition, however, and the scores in that condition were significantly higher. What we would need then is some attribute of the participant that reduces learning in the distributional-cue-only situation more than in the distributional-cue-plus-
lengthening situation. If statistical calculations require sustained attention in a way that lengthening does not, this account could explain the differences without reference to a competing rhythmic cue. If such an account is not plausible, we would be left with a competing, noisy cue in the language.

In Chapter 5, we turn to Experiment 2, which uses a mixed disyllabic and trisyllabic language.
CHAPTER 5: EXPERIMENT 2 -- LANGUAGE WITH DISYLLABIC AND TRISYLLABIC WORDS AND VARYING WORD DURATIONS

5.1. Methodology

Experiment 2 follows a similar methodology to Experiment 1. The only modification is that the artificial language was composed of 3 disyllabic l-words and 3 trisyllabic l-words, instead of all disyllabic words as in Experiment 1. This experiment extends the results of Tyler & Johnson (2006) who found that infants could not learn a language such as this. The participants in this study are all adults.

5.1.1. Participants.

Thirty-six adults participated in Experiment 2, 18 in Language A and 18 in Language B. No one who participated in Experiment 1 participated in Experiment 2.

5.1.2. Stimuli.

Language A was composed of three disyllabic words and three trisyllabic words. The specific l-words for Language A were: *digu, kuba, bupi, kaduda, gakitu, and tipabi*. These l-words are based on the words of Experiment 1 with a new syllable added for the trisyllabic l-words. This was done since no bias for Language A or B was found in the results for Experiment 1. Boundary words were *gukadu, badi, daku, bibu, pigaki, and tutipa*. L-words and boundary words are found in Table 4.

<table>
<thead>
<tr>
<th>Table 4. L-Words and Boundary Words for Experiment 2.</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-Words</td>
</tr>
<tr>
<td>Boundary Words</td>
</tr>
</tbody>
</table>

All syllables were 330 ms in duration, creating disyllabic l-words of 660 ms and trisyllabic l-words of 990 ms. Since some words were now 33% longer than others, a
pure 7-minute block was not possible for a direct comparison to Experiment 1. Since Tyler & Johnson’s (2006) results suggested a failure to learn was likely, it was decided to keep the same number of l-word tokens as Experiment 1, meaning the same number of tokens are heard in both experiments, with the resulting training block for Experiment 2 being approximately 8 and a half minutes in length. During testing, all l-words and boundary words were paired exhaustively for 36 test pairs. Language B used Language A’s boundary words as l-words and an identical test.

5.1.3. Procedures

Experiment 2 followed the same procedures as Experiment 1. The total time to conduct a session was slightly longer, about 11 minutes, due to the longer training block.

5.2. Results: Analysis of Means

Thirty-six people participated in Experiment 2, 18 in Experiment 2A and 18 in Experiment 2B. During analysis, one participant’s results were excluded from each language, 2A and 2B, due to being more than 2 standard deviations from the mean. Therefore, N=17 in both 2A and 2B. Both languages scored above chance (2A: M=0.5523, SD=0.0528; 2B: M=0.5523, SD=0.0680). Figure 6 displays means and confidence intervals for both experiments.

Again, the normality of the distributions was assessed visually using a plot of the density curve for both 2A and 2B (Figure 7). A Shapiro-Wilk test for normality was then conducted, indicating that neither 2A nor 2B significantly violated normality (2A: W=0.9567, p=0.57; 2B: W=0.9508, p=0.4699).
Figure 6. Mean proportion correct with confidence intervals for Experiment 2.

Figure 7. Density curves for Experiments 2A and 2B.
The variances for 2A and 2B were not significantly different \( (F(16,16)=0.6032, p=0.3221) \). With both equal variance and a normal distribution, we can use a one-way ANOVA to assess the means of 2A and 2B. The means were not significantly different \( (F(1,32)=4.94\times10^{-30}, p=1) \) between the two languages. As the means were identical beyond 4 decimal places, it is not surprising that there is no evidence for differing means for the two languages.

Experiments 2A and 2B were then collapsed together for further analysis. A Shapiro-Wilk test confirmed that the consolidated results followed normality \( (W=0.9603, p=0.2475) \). The mean for Experiment 2 \( (M=0.5523, SD=0.5992) \) is slightly below that of Experiment 1 \( (M=0.5903, SD=0.1344) \). However, due to increased power and reduced variance, a one-sample t-test demonstrated that the mean for Experiment 2 was still significantly different from chance \( (t(33)=5.0886, p<.0001) \).

![Exp 1 / Exp 2 Mean plus Confidence Interval](image)

Figure 8. Mean proportion correct with confidence intervals for Experiments 1 and 2.
As the mean for Experiment 2 (.55) is slightly under that of Experiment 1 (.59), it would be interesting to see if the difference is significant. First, a parametric test of variance was performed. The variances are clearly not the same ($F(23,33)=5.0325$, $p<.0001$). Due to this violation of one of the assumptions of the one-way ANOVA, the non-parametric Kruskal-Wallis test was used to compare Experiment 1 and Experiment 2. While there is a difference, it does not yet reach significance ($\chi^2(1)=2.1805$, $p=0.1398$). Means and confidence intervals are plotted in Figure 8.

5.3. Results - Analysis of Selections

In this section, we will attempt to reconstruct the participant’s behavior during training and testing as much as the data will allow. The method will be to posit the possible ways a participant might be segmenting the speech during training, and then compare those models with their word selections during the testing phase. The two principles below inform the proposal given:

1) Word segmentation will occur immediately and continuously.
2) The participant will attempt to apply their phonological system to the language at all times.

Looking first at principle 1, syllabic transitional probabilities (SylProbs) provide perfect markers to word boundaries in all artificial languages used during this set of experiments. The SylProbs are always 100% within an l-word and 20% between l-words. This remains true no matter what testing language is used and what durations are used. However, this statistical pattern only unfolds over time as the data arrives. *The participant will not wait until the time series has unfolded to segment the speech.* They will segment immediately with whatever cues are available. This implies that any learning of the statistical word boundaries will be done in the context of an already segmented speech stream (see
Section 1.3.1 and 3.3.1 for previous discussion of this). Principle 2 simply states that participants will hear all speech using their language’s phonological grammar.

5.3.1. A proposal for Experiment 2A

We will want to start by looking for any cues that are immediately available for segmentation, no matter how fragmentary. Experiment 2 was designed so that there would be no cues for the words other than statistical patterns. Experiment 2’s words were formed by adding a final syllable to three of Experiment 1’s words, and Experiment 1 showed no bias for Language A or B. However, as mentioned in Section 4.1.2, aspiration of voiceless stops occurs at the beginning of a foot in English but does not occur in other contexts. For this reason, voiceless stops were distributed in different syllables so that they would not reliably signal a word onset. It is possible that participants did use this allophonic cue to segment speech even though it was not a reliable cue for l-words. An aspiration cue will be available immediately as soon as the language begins.

To look for evidence of this, we will need to look in detail at the specific words of the language and their selection during the test. Example (28) contains a sample of language A with the voiceless stops highlighted in bold. Since English primarily has binary feet composed of syllables, (29) contains the resulting feet (in italics) that would be heard using the aspiration segmentation strategy. Note that voiceless stops sometimes occur in adjacent syllables. In such cases, we can posit that both pairs of syllables might be heard as feet.

(28)  \textit{tipabikubabupikadudadigugakitu} \\
(29)  \textit{tipabikubabupikadu}dadigugakitu
The resulting feet created from using Initial Voiceless stops as the starting point of a metrical foot (henceforth IV-feet) are listed in (30).

(30)  **tipa, pabi, kuba, pika, kadu, kitu**

Of course, (30) is based on a sample. When looking at the entire language, we quickly see that 4 of these IV-feet behave differently than all the other ones. Namely, 4 IV-feet occur word internally, while the others cross a word boundary. The word-internal IV feet are presented in bold in (30). Three of those word internal feet occur within a trisyllabic word, while one is a complete disyllabic word from language A. All other possible IV-feet that occur across a word boundary will be heard, but heard less consistently as they change form with each sequence of l-words. The l-words were concatenated pseudo-randomly for the training materials. Now, let us compare these repeatedly heard IV-feet to the l-words and boundary words of language A, presented in (31) and (32), respectively.

(31)  **tipabi, kaduda, gakitu, bupi, digu, kuba** (l-words)

(32)  **tutipa, gukadu, pigaki, bibu, badi, daku** (boundary words)

The bold font in (31) and (32) highlights the presence of the IV-foot inside the l-words or boundary words. If participants based their test selections simply on words that contained the IV-feet, then the l-words and boundary words in bold should be the most selected items. That list is: **tipabi, tutipa, kaduda, gukadu, gakitu, and kuba**. For instance, the IV-foot **tipa** is included within the l-word **tipabi** and the boundary word **tutipa**. Similarly, the IV-foot **kadu** is included within the l-word **kaduda** and the boundary word **gukadu**. We would expect that participants would have a difficult time telling which of these words is the correct one if the IV-foot is their only information. On the other hand,
the IV-foot *kuba* is an l-word and has no competition from boundary words. We would therefore expect *kuba* to be selected frequently.

Table 5 contains both a table of selected test words ordered by proportion selected, plus a figure displaying the same information. Within the figure, *L* and *B* mark l-words and boundary words respectively. During testing, all six l-words and boundary words are presented 6 times each and fully crossed, making for 36 test items. A chance selection of any word would be 3 of 36 items or 0.0833. As Table 5 shows, only 5 words are selected over this 0.0833 threshold. Those words, in order, are: *kaduda, tipabi, tutipa, gukadu, and kuba*. All 5 of these words are the words predicted from the IV-list. The only exception is that *gakitu*, supported by the IV-foot *kitu*, occurs in the middle of the figure.

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>kaduda</td>
<td>0.127</td>
</tr>
<tr>
<td>tipabi</td>
<td>0.109</td>
</tr>
<tr>
<td>tutipa</td>
<td>0.102</td>
</tr>
<tr>
<td>gukadu</td>
<td>0.096</td>
</tr>
<tr>
<td>kuba</td>
<td>0.086</td>
</tr>
<tr>
<td>bupi</td>
<td>0.080</td>
</tr>
<tr>
<td>digu</td>
<td>0.073</td>
</tr>
<tr>
<td>gakitu</td>
<td>0.071</td>
</tr>
<tr>
<td>pigaki</td>
<td>0.068</td>
</tr>
<tr>
<td>bibu</td>
<td>0.066</td>
</tr>
<tr>
<td>badi</td>
<td>0.060</td>
</tr>
<tr>
<td>daku</td>
<td>0.055</td>
</tr>
</tbody>
</table>

A linear mixed effects model was constructed to analyze the significance of voiceless stops in the words that a participant selects. In the model, a binary Accuracy
measurement served as the dependent variable. Four binary predictors were coded to mark the presence of a predictor on each individual test item: 1) Voiceless stop on an l-word’s first syllable, 2) voiceless stop on an l-word’s second syllable, 3) voiceless stop on a boundary word’s first syllable, and 4) voiceless stop on a boundary word’s second syllable. If a voiceless stop did occur in the onset to the syllable, it was coded as T, while an absence was coded as F. The third syllables were not coded since half of the l-words were disyllabic. Finally, the participant and the interaction of word and boundary word were coded as random effects. This was done since the words are a sample of all the possible words that might have been used, and the selection of the word interacted with its test foil, the boundary word.

Predictors 1, 2, and 4 were all significant as displayed in Table 6. If the 1st syllable of a word contained a voiceless stop, the participant was more likely to choose it correctly. However, if either the 1st or 2nd syllable of a boundary word contained a voiceless stop, they were significantly more likely to select it, incorrectly.

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Wald’s Z</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>l-word 1st syllable</td>
<td>0.8302</td>
<td>4.214</td>
<td>&lt;.0001*</td>
</tr>
<tr>
<td>l-word 2nd syllable</td>
<td>-0.1357</td>
<td>-0.690</td>
<td>0.4900</td>
</tr>
<tr>
<td>boundary word 1st syllable</td>
<td>-0.4450</td>
<td>-2.384</td>
<td>0.0171*</td>
</tr>
<tr>
<td>boundary word 2nd syllable</td>
<td>-0.4120</td>
<td>-2.318</td>
<td>0.0204*</td>
</tr>
</tbody>
</table>

Simply dividing the speech stream by IV-feet is sufficient to show moderate learning. If a participant selected the word *kuba* each time, since it is an IV-foot in itself, and then performed at chance on all other words, this would sum to 21 correct selections. This would generate a proportion correct of 0.5833, which is very similar to the proportions we saw in most experiments. The results for 2A were in fact slightly lower than that.
The pattern of selected words in Table 5 does not support this view, however. First, we see that the four most selected words are all tri-syllabic. The IV-feet-only strategy gives no reason to expect this. If this was just one experiment, this might have happened by chance, but, as we will see, tri-syllabic words are consistently preferred throughout the set of experiments. Moreover, if words are selected entirely on the basis of whether they contain an IV-foot, then the words not containing an IV-foot should be chosen at chance. This is not what we see.

All three disyllabic l-words are chosen more often than all three disyllabic boundary words. As predicted by the IV-foot theory, kuba is selected the most of all disyllabic l-words, but the IV-foot theory does not predict the correct selection of the other disyllabic l-words, since none of them contain IV-feet. Also, why is kuba not selected more than all other words, since it is an exact match to the IV-foot, and the trisyllabic words contain an IV-foot but have an additional syllable? We will return to this, but there is evidence for a trisyllabic bias in participant selections. For now, let us grant the possibility of a trisyllabic bias and investigate its presence.

Since the 3 disyllabic and 3 trisyllabic l-words were crossed fully to generate the test items, this means that 9 of the test items are a trisyllabic-trisyllabic pair, 9 are a disyllabic-disyllabic pair, and 18 are a trisyllabic-disyllabic pair. In the test items, the order of a disyllabic test word and a trisyllabic test word are balanced so that one type is not always first.

17 We can then investigate how disyllabic words were selected when a trisyllabic word is not a possible choice. Table 7 presents the selection of test words grouped by the type of word pairs (disyllabic-disyllabic, trisyllabic-disyllabic, and trisyllabic-trisyllabic).
As can be seen, when only disyllabic-disyllabic pairings are considered, all three l-words were selected more often than all three boundary words. However, when considering trisyllabic-disyllabic pairings, all six trisyllabic words are selected more than all six disyllabic words. When a trisyllabic option is not present, as a group, participants know that three of the disyllabic words are the correct ones, but when they hear a disyllabic word paired with a trisyllabic word, even with a trisyllabic boundary word, they have a tendency to select the trisyllabic word. Within these trisyllabic-disyllabic pairings, we still see the disyllabic l-words being selected more often than disyllabic boundary words, but the trisyllabic words are jumbled together, with some boundary words being selected more than l-words. This confusion among trisyllabic words remains even when only trisyllabic-trisyllabic pairs are considered. The least selected word in the trisyllabic-trisyllabic pairs is an l-word.

5.3.2. A Bias for Trisyllabic Words

Why should there be a bias for trisyllabic words? One possibility worth considering is that the experiment itself does not induce the trisyllabic bias, but that the
participants bring the bias to the experiment. This was investigated through a brief experiment in which the testing materials for Experiments 2 and 3 were presented to 10 participants without any training phase. This should reveal any bias for words or boundary words inherent in the testing materials. Recall that for languages A and B in all experiments, the testing materials are identical. It is only the training language that distinguishes an A and a B experiment. However, the testing materials are different when moving from Experiment 2 to Experiment 3 (Section 6.1 provides the Methods for Experiment 3). In Experiment 2, all syllables were of the same duration, 330 ms, so the testing words also all have syllable durations of 330 ms. In Experiment 3, trisyllabic words have shorter syllable durations than disyllabic words. So as not to bias test words in either direction, testing materials in Experiment 3 were presented with syllables having a duration that was the mean of the disyllabic syllables and the trisyllabic syllables. This mean was 245 ms. (see Section 6.1 for details).

In short, in this new experiment, the test words of Experiments 2 and 3 with syllable durations of either 330 ms or 245 ms, respectively, were presented to 10 participants. Five participants listened to each set of 36 test items. All were asked to mark which word of the pair they preferred with no guidance provided as to what “preferred” might mean. Table 8 presents the number of times the participants selected a trisyllabic or disyllabic word in the two conditions only for the trisyllabic-disyllabic pairs. When the participant was offered the choice of selecting either a trisyllabic or disyllabic word, did they have a preference?

<table>
<thead>
<tr>
<th></th>
<th>245 ms duration</th>
<th>330 ms duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>trisyllabic words</td>
<td>81 (75%)</td>
<td>45 (41.67%)</td>
</tr>
<tr>
<td>disyllabic words</td>
<td>27 (25%)</td>
<td>63 (58.33%)</td>
</tr>
</tbody>
</table>

Table 8. Counts of Word Selection Without Training
In the 330 ms condition, participants had no bias for a trisyllabic word (58% of selections were for disyllabic words), but in the 245 ms condition, they did have a bias (75% of selections were for trisyllabic words). The latter is clearly of interest in interpreting the results of Experiment 3, as that is the condition in which test words had syllable durations of 245 ms. However, for the current examination of Experiment 2 where syllable durations are 330 ms, this brief test provided no evidence that participants came to the experiment with a bias for trisyllabic words. This suggests that, for Experiment 2, any bias for trisyllabic words arises from the training process.

We can also look for any bias that the particular words in Languages A and B might have. Again, considering only trisyllabic-disyllabic pairs, Table 9 presents the number of times participants chose words from language A or from Language B. The participants do choose the Language B words more often, but the difference is quite small.\(^{18}\)

<table>
<thead>
<tr>
<th></th>
<th>\underline{245 ms syllable duration}</th>
<th>\underline{330 ms syllable duration}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Language A</td>
<td>Language B</td>
</tr>
<tr>
<td>Disyllabic Words</td>
<td>12 (44%)</td>
<td>15 (56%)</td>
</tr>
<tr>
<td>Trisyllabic Words</td>
<td>39 (48%)</td>
<td>42 (52%)</td>
</tr>
</tbody>
</table>

5.3.3. A trisyllabic bias through rhythmic perception

Why might participants be choosing the trisyllabic words more often? We previously noted that the IV-feet within trisyllabic l-words are more stable than those crossing a word boundary in that they consistently recur. The same feature is seen if we reconsider the timing pattern of Experiment 2. The syllable durations of Experiment 2 are uniform throughout training. As discussed in Section 3.1.2, this is likely to induce

\(^{18}\) Unfortunately, the sample size is too small for a meaningful statistical analysis.
rhythmic attention towards the language, and both music psychology and metrical phonology predict that the tendency will be for the participant to divide the syllables into groups of two. If the listener has no indication where to start the rhythmic parsing, they may start at any point in the speech. For disyllabic l-words, this randomly placed metrical grouping will sometimes happen to coincide with a word and sometimes will divide the word in half. However, no matter where the rhythmic grouping is placed, a trisyllabic l-word will always have two of its syllables grouped together.

Examples (33) and (34) contain a sample of the training language A. In each, the alternation of bold and regular fonts indicates syllable pairs. In (33), the grouping starts on the first syllable, while, in (34), the grouping starts on the second syllable. Trisyllabic l-words are underlined. Note that no matter which binary parse is used, (33) or (34), the underlined trisyllabic words always have two syllables grouped together.

(33)  \textbf{tipabikubabupikadudadigugakitu}

(34)  \textbf{tipabikuba}bupikadudadigugakitu

Considering the disyllabic l-words again, let us take the l-word \textit{digu}. A rhythmic grouping that happens to include the l-word \textit{digu} will have found \textit{di} paired with \textit{gu}, but other rhythmic groupings will split the l-word \textit{digu} such that \textit{di} is paired with 5 other words. Without considering statistics, all 6 pairs containing the syllable \textit{di} would be viable word choices. Now, we will look at the trisyllabic word \textit{gakitu}. Some rhythmic parses will pair the middle \textit{ki} syllable with \textit{ga} and sometimes with \textit{tu}, but that is the end of the variability for \textit{gakitu}. The listener might not be able to tell the difference between a \textit{gaki} pairing and a \textit{kitu} pairing, but those are the only two options they will have heard for \textit{gakitu} (only considering a random rhythmic mapping).
As such, it seems reasonable that those syllable pairings, found within trisyllabic words, would seem more familiar than the syllables within disyllabic words. The claim is not that rhythm uniquely explains the tendency in Experiment 2 to choose trisyllabic words. A mapping of metrical feet based on the beat of the language simply contributes to this, as does the repetition of IV feet within trisyllabic words. This explanation based upon random rhythmic groupings of two syllables does predict that an l-word and boundary word will cluster together in proportion selected. When examining the trisyllabic-trisyllabic pairings in Table 7, tipabi and tutipa are found next to each other, as are pigaki and gakitu. Only the kaduda and gukadu pair does not match the prediction.

5.3.4. Statistical Feet

At this point, we have made headway in explaining the words that participants select in Experiment 2A, but we do not have an adequate theory yet. The main thing we cannot yet explain is why the participants correctly choose the right disyllabic l-words, while being somewhat confused about which trisyllabic words are correct.19 When we look at the set of results for all the experiments, SylProbs (transitional probabilities between syllables) might not seem promising. After all, dips in transitional probabilities uniquely mark all words in all languages, but participants do not behave equally successfully in all experiments. That would appear to suggest that SylProbs are not being used. I will argue against that suggestion.

The key idea is to consider the possibility that participants are only remembering pairs of syllables, instead of the entire word, from statistical grouping. Of course, with a disyllabic l-word, if a pair of syllables is remembered, then the entire word is

19 The rhythmic grouping notion might explain the trisyllabic confusion, but not the success at disyllabic words.
remembered. This is untrue for trisyllabic l-words. As discussed in Section 1.5, it is possible that a change in transitional probability is understood as a prosodic boundary. This was based upon the ubiquitous connection between prosodic boundaries and word segmentation as well as the new proposal that a sudden rise in uncertainty will be experienced as a slight durational lengthening (Section 1.4).

If the dip in transitional probabilities marks the end of a prosodic unit, it is logical for that to be the disyllabic foot, as all content words in English consist of a prosodic foot (though degenerate feet are possible). We might term these syllable pairs Statistical Feet. This is not a new type of metrical foot. A statistical foot is simply a metrical foot whose boundary was best cued by a statistical pattern, just as an IV-foot is a metrical foot located using the aspiration of a voiceless stop. Example (35) presents how the participant would hear a speech sample with statistical feet.

(35)  **tipabi kuba bupi kaduda digu gakitu**

Example (35) presents the same speech sample of language A as seen in (33) and (34) above. In this case, spaces have been added at each point at which the SyllProbs dip from 1 to 0.2. These statistical boundaries mark the l-words completely as that was the design of the language. The bold font highlights the last two syllables of each l-word. For disyllabic l-words, the statistical foot is the entire l-word, while for trisyllabic l-words the first syllable is left either unparsed or separate from the foot. The resulting statistical feet are listed in (36).

(36)  pabi, kuba, bupi, duda, digu, kitu.

Three of these words are the disyllabic l-words in their entirety. Statistical feet then provide an explanation for why participants in 2A were able to choose the correct
disyllabic l-words at least in disyllabic-disyllabic pairs. The other three statistical feet are, of course, components of the trisyllabic l-words, so we have an additional reason for participants to choose them. There remains a question, however. If we compare these statistical feet to the trisyllabic boundary words, the statistical feet are not components of them. The trisyllabic boundary words for 2A were *tutipa, gukadu*, and *pigaki*. Each of the boundary words only has a single syllable overlap with the feet of (36). This implies that statistical feet should help participants choose only the correct l-words and not make errors on the boundary words. Why is that not the case?

The fact that IV-feet are available immediately for segmentation, while statistical feet are only available after the statistical structure has unfolded through time, could explain this. By the time the participant has started to hear statistical feet, they already have a collection of IV-feet as possible words. All of the trisyllabic l-words then have already been parsed and the new statistical feet must compete with that previous parse. This problem could impact the disyllabic words as well, but it should be less than for trisyllabic l-words, since IV-feet are more stable within the trisyllabic words (and in *kuba*). This recalls the garden pathing effect that Gebhart et al. (2009) documented in which it took tripling exposure to a second language to allow a participant to back away from the structure of a first language.

As an example, we may consider the disyllabic l-word *bupi*. Using voiceless stops as the cue for feet, a participant may have taken *pi* as the start of a foot. However, the second part of the foot will be unstable since it crosses an l-word boundary. The participant might have parsed *piti, pika, piga, pidi, or piku*, since all of those combinations will occur in the randomly concatenated language stream. Compare this to
*kadu* inside *kaduda*. Every single time the l-word *kaduda* occurs in the speech stream, *ka* is paired with *du*. This implies a much stronger memory trace for *kadu* than the randomly changing *pi*-X segmentations. Unless the participant has a way to drop *kadu* from their segmentation, they could be confused by the boundary word *gukadu*.

Neither statistical feet, rhythmic groupings, nor IV-feet alone are sufficient to explain the results of Experiment 2A. Statistical feet suggest greater success than in fact occurs, while IV-feet have no way to account for some of the genuine success, such as with the disyllabic l-words. However, if we assume that the participant will start to parse immediately and continuously, they will first grab IV-feet and then supplement that with statistical feet. This most closely matches the results as seen in Table 5.

### 5.3.5. Statistical feet in Saffran et al. (1996)

The original Saffran et al. (1996) paper contained several different tests. One test pitted trisyllabic l-words against trisyllabic non-words, which are foils composed of non-adjacent syllables from the artificial language. Learning in that condition was at 76%. A second test pitted the trisyllabic l-words against trisyllabic part-words, which are foils composed of two syllables of an l-word and one other syllable from the language. In this case, the learning achieved was at 65%. Participants, however, could not distinguish all test items equally well in the part-word case.

Half of the part-words in Saffran et al. (1996) were composed of the first two syllables of an l-word and the other half were composed of the second two syllables of an l-word. When the part-word foils contained the first two syllables of an l-word, participants in Saffran et al. (1996) only selected the foils 29% of the time. However, when the foil contained the second two syllables, they performed at chance. In other
words, the participants could not tell the difference between an l-word and a part-word containing the last two syllables of an l-word. This makes sense if they are only learning the last two syllables of an l-word, i.e., if they are only learning a statistical foot. Saffran et al. suggest that this discrepancy occurs because the participants are learning the morphological word but are better able to remember the ends of words than the beginning. This is a viable explanation and matches previous research on various memory tasks where the final edge of an event or a list is recalled more accurately than other components (see this data and discussion in Saffran et al [1996, pp. 614-615]), but one might interpret these results prosodically instead.

5.3.6. Putting it all together for Experiment 2A

To review, we have employed three factors to explain the results of Experiment 2A. As soon as the language is encountered, the participant will try to segment using their phonological knowledge. Two pieces of information are available to them within the first words – the regular rhythmic beat and the aspiration of voiceless stops. The participants will search for structure in the language with these tools (Gomez, 2006; Onnis et al., 2004). They will select certain feet (IV-feet) based upon the occurrence of aspirated voiceless stops. Since every syllable has an equal duration, they will also try to find a connection between the words and the rhythm. However, due to the varying word lengths and the random concatenation used to create the language, there is no way, at the beginning, to find a rhythm that continuously groups the same syllables. However, no matter how participants lay the rhythmic groupings upon the syllable stream, at least two syllables within the trisyllabic words will be heard, though which two syllables will vary.
Participants will also continuously try to predict the next syllable in the same way that listeners try to predict the syntactic or lexical structure of the speech they hear. As the language continues and the statistics emerge, they will find that they can predict some syllables, but others remain difficult. This increase in uncertainty will be experienced as a slight lengthening compared to other syllables, which cues the end of a prosodic foot. Participants can now hear these statistical feet reoccurring throughout the speech and they must somehow unify this information with their other perceptions.

Unfortunately, we do not know how this unification is achieved. One possibility is that there is no unification. The listener simply has memories of all the different feet they have heard. This collection of memory traces might then be used during the test.

Another possibility is that there are attempts to put the groupings together. For instance, a complete solution would take the IV-foot *kadu* and the statistical foot *duda*, see the possible connection, and then put them together into a single word *kaduda*. If participants do this, then the complete structure of the language will fall out. Interestingly, we have no evidence that any participant did this. The highest score was 64% for any individual. Why was the complete structure not found? One possibility is the confusing rhythm. Because of the identical syllable durations, participants will continue to use what they have learned and predict the next group of syllables. When, for instance, 2 or 3 disyllabic words happen to occur in a sequence, it will seem as if the rhythm is finding structure in the language, but then a trisyllabic word will occur and a false word will have been heard.

To make this clear, let us put the l-words *bupi, digu, kuba, kaduda,* and *bupi* in a sequence.
The participant may have heard *bupi* before and therefore starts another rhythmic grouping after it (cf. Dilley & MacAuley, 2008), which creates the following sequence of syllable pairs:

(38) bupi digu kuba kadu dabu pi—

The first three pairs are complete l-words, but the next three are not. These rhythmic mappings will continue to break up words and prevent the full structure from being found. The overall experience then will remain, not one of hearing a complete parsing of everything into words, but the repeated recurrence of words within noise that cannot be adequately segmented.

### 5.3.7 Why do adults behave differently than children?

In Experiment 2, adults chose words significantly over chance. The learning was quite weak, 55%, and a detailed analysis revealed that only disyllabic l-words were selected over disyllabic boundary words consistently, but nevertheless the learning was significant. However, the 8-month-old infants in Tyler & Johnson (2006) did not show any ability to discriminate l-words and boundary words in a similar language. It is difficult to compare directly because the infants in Tyler & Johnson were acquiring Dutch, while the adults in Experiment 2 were native English speakers. The participants in Experiment 2 appear to have used some language-specific knowledge about the correlation of aspiration and metrical structure to segment the language, the so-called IV feet. Infants may not have access to this sort of knowledge at 8 months. The general trend in phonological acquisition is that infants can make a broad range of phonetic discriminations before their first birthday, even discriminations irrelevant to the language.
they are acquiring. The 8-month-olds should be able to hear the various allophones of voiceless stops in the artificial language. However, it is only around a year old that they start to display phonological knowledge specific to their language, such that they begin to stop making discriminations among sounds that are unimportant in the language they are learning (Kuhl, 1992; Maye et al., 2002; Werker & Tees, 1984). While the infants would hear cues such as aspiration, they likely do not know its phonological significance yet. Of course, Dutch will have different allophonic patterns than English, so Tyler and Johnson (2006) must be reassessed in terms of Dutch phonological structure.

Infants do, however, have a strong sense of rhythm before one year. 2- to 3-month olds can distinguish isochronous sequences from "rhythmic sequences," which have strong and weak patterns (Demany, McKenzie, & Vurpillot, 1977). 5-month olds can differentiate different rhythmic groupings (Chang & Trehub, 1977). More recently, Phillips-Silver and Trainor (2005) show that infants of 7 months can discriminate ambiguous rhythms with a beat, perceiving different patterns from the same stimulus based upon when they are bounced by a care-giver to mark prominent beats. This is relevant because it shows that 7-month-olds perceptually impose rhythm on the stimulus, as opposed to deriving it purely externally. Finally, infants of 9 months can track rhythms where the tones are precisely at the 300 ms durations of natural speech (Bergeson & Trehub, 2006)²⁰.

It is precisely at the rhythmic level, however, where Experiment 2 is uninformative for finding words. Syllabic onsets occur at a regular rhythm but word

---

²⁰ We have no evidence that infants at 7-months require bodily movement to perceive rhythm at this rate. The repeated ability of infants to make various rhythmic discriminations both in music and speech suggests they do not require bodily movement.
onsets do not. Since the rhythmic pattern is available early, before statistics can emerge, infants may be attempting to use rhythmic groupings and getting garden-pathed away from the correct structure. It is worth remembering, however, that all of the extra phonological knowledge adults have only let them score at 55%.

A final difference between Tyler and Johnson’s research and the work with adults here is that infants only went through 2 minutes of training, while the adults here went through 8.5 minutes of training. Possibly, additional training for either group would increase accurate segmentation.

5.3.8. **Summary of Experiment 2A**

The results of Experiment 2A are most compatible with the Rhythm And Statistics hypothesis. We could not explain the word selections without recourse to both. If only statistical learning was being used, the participants should have done far better than they did. The lack of a match between the rhythmic pattern and the words onsets, in turn, is the most likely reason learning was not better. Experiment 2A also emphasizes the importance of considering the time course of learning. Cues such as a rhythmic pattern or IV-feet that are immediately available will compete with the statistical cue that emerges over time, whether or not one is “weightier” than another.

5.4. **Applying the Framework to Experiments 2B.**

We now need to see if this framework of overlapping feet – IV feet, statistical feet, and rhythmic feet – accurately explains the results of Experiment 2B. The mean proportion correct for 2B was identical to 2A, 55%. As before, voiceless stops are immediately available as possible IV-feet markers. The l-words of 2B are the boundary words of 2A and that list is: *lutipa, gukadu, pigaki, badu, bibu*, and *daku*. 
If we focus only on the feet that regularly occur within an l-word, since those are the ones that recur without change throughout the training, we end up with 4 IV-feet: *tuti*, *tipa*, *kadu*, and *piga*. These IV-feet occur within all tri-syllabic l-words from which they were drawn, as well as within the boundary words *tipabi* and *kaduda*. Let us compare this to the actual participant selections.

Table 10. Experiment 2B Selected Words

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>gukadu</td>
<td>0.123</td>
</tr>
<tr>
<td>gakitu</td>
<td>0.101</td>
</tr>
<tr>
<td>tipabi</td>
<td>0.100</td>
</tr>
<tr>
<td>tutipa</td>
<td>0.098</td>
</tr>
<tr>
<td>pigaki</td>
<td>0.095</td>
</tr>
<tr>
<td>badi</td>
<td>0.090</td>
</tr>
<tr>
<td>kaduda</td>
<td>0.085</td>
</tr>
<tr>
<td>daku</td>
<td>0.083</td>
</tr>
<tr>
<td>bupi</td>
<td>0.069</td>
</tr>
<tr>
<td>bibu</td>
<td>0.064</td>
</tr>
<tr>
<td>kuba</td>
<td>0.051</td>
</tr>
<tr>
<td>digu</td>
<td>0.042</td>
</tr>
</tbody>
</table>

As with 2A, all 5 words that contain IV feet are selected above the chance 0.0833 threshold. In particular, the participants do not seem to be able to tell the difference between *tutipa* and *tipabi*, which both contain the IV-foot *tipa*. A mixed effects model using voiceless stops as predictors confirms statistical significance (Table 11).

Table 11. Mixed Effects Model of Voiceless Stops as Predictors of Accuracy

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Wald’s Z</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>l-word</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st syllable</td>
<td>0.6988</td>
<td>2.344</td>
<td>0.0190*</td>
</tr>
<tr>
<td>2nd syllable</td>
<td>-0.0565</td>
<td>-0.206</td>
<td>0.8365</td>
</tr>
<tr>
<td>boundary word</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st syllable</td>
<td>-0.6382</td>
<td>-2.345</td>
<td>0.0190*</td>
</tr>
<tr>
<td>2nd syllable</td>
<td>-0.8658</td>
<td>-3.072</td>
<td>0.0021*</td>
</tr>
</tbody>
</table>
We should also look at the selections by syllable type pairing (Table 12). Similar to 2A, the 5 most popular selections are all trisyllabic words, despite the fact that two of them are boundary words (Table 10). When only disyllabic words are available for selection, all three disyllabic l-words are selected over all three disyllabic boundary words. However, if a disyllabic word is paired with a trisyllabic word, they abandon the disyllabic words such that only trisyllabic words are chosen above chance. Finally, when only trisyllabic words are available for selection, the participants show little ability to distinguish them. These are the same patterns seen in 2A.

Table 12. Experiment 2B Selected Words Grouped by Word Pairings

<table>
<thead>
<tr>
<th>Disyllabic-Disyllabic</th>
<th>Trisyllabic-Disyllabic</th>
<th>Trisyllabic-Trisyllabic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>Frequency</td>
<td>Word</td>
</tr>
<tr>
<td>badi (L)</td>
<td>0.222</td>
<td>pigaki (L)</td>
</tr>
<tr>
<td>daku (L)</td>
<td>0.203</td>
<td>gakitu(B)</td>
</tr>
<tr>
<td>bibu(L)</td>
<td>0.190</td>
<td>gukadu (L)</td>
</tr>
<tr>
<td>bupi (B)</td>
<td>0.163</td>
<td>tutipa (L)</td>
</tr>
<tr>
<td>kuba (B)</td>
<td>0.131</td>
<td>tipabi (B)</td>
</tr>
<tr>
<td>digu (B)</td>
<td>0.092</td>
<td>kaduda (B)</td>
</tr>
<tr>
<td>badi (L)</td>
<td>0.069</td>
<td></td>
</tr>
<tr>
<td>daku (L)</td>
<td>0.065</td>
<td></td>
</tr>
<tr>
<td>bupi (B)</td>
<td>0.056</td>
<td></td>
</tr>
<tr>
<td>digu (B)</td>
<td>0.039</td>
<td></td>
</tr>
<tr>
<td>kuba (B)</td>
<td>0.036</td>
<td></td>
</tr>
<tr>
<td>bibu (L)</td>
<td>0.033</td>
<td></td>
</tr>
</tbody>
</table>

As in 2A, the words containing IV feet are frequently selected, but we need more than IV-feet to explain the results. For instance, none of the disyllabic words are made up of IV-feet, but, within the disyllabic-disyllabic pairings, participants reliably choose the disyllabic l-words, not the boundary words. As in 2A, statistical feet can provide us with a motivation for this. The statistical feet for 2B are marked in (39) using bold.

(39) **tutipa daku bibu gukadu badi pigaki**

This yields the statistical feet in (40).
(40)  daku bibu badi tipa kadu gaki

The first three statistical feet are disyllabic l-words. 2A and 2B are slightly different with regard to the latter three statistical feet that are within trisyllabic l-words, however.

The trisyllabic boundary words for 2A were created by joining the final syllable of one word and the first two syllables of another word. The language of 2B is simply the boundary words from 2A with 2A’s l-words as the boundary words. This implies that the boundary words for 2B contain the last 2 syllables of an l-word, while the boundary words of 2A contain the first 2 syllables of an l-word. This is displayed in Table 13. The boundary words for 2A contain A&B indices, while the boundary words for 2B contain B&C indices.

Table 13. Words and Boundary Words for Experiments 2A and 2B

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>2A L-Words</td>
<td>d2Agu1B</td>
<td>k23Ba2B</td>
<td>b24P13B</td>
<td>k4Adu4gd4Ac</td>
<td>g4Aku4g1u4Sc</td>
<td>u4Apa4g1b4Sc</td>
</tr>
<tr>
<td>2A B-Words</td>
<td>gu1Bka4Adu4B</td>
<td>b23Adi1A</td>
<td>d4ak23A</td>
<td>b4cbu3A</td>
<td>p13Ga3gk13B</td>
<td>u4ctu4Apa4B</td>
</tr>
<tr>
<td>2B L-Words</td>
<td>gu1Bka13du1C</td>
<td>b23Ad12B</td>
<td>d3Aku13B</td>
<td>b1Apu4B</td>
<td>p13Ga3gk13C</td>
<td>u4ctu4Apa4C</td>
</tr>
<tr>
<td>2B B-Words</td>
<td>d12Bg1u1A</td>
<td>k3Bba2A</td>
<td>b4Ap13A</td>
<td>k14du1Cd13A</td>
<td>g13Ak15k16A</td>
<td>u16Ap13cb14A</td>
</tr>
</tbody>
</table>

Since the test boundary words for 2B are made of the second two syllables of the word, instead of the first two, the statistical feet occur in all three boundary words and all three l-words. Moreover, the statistical feet and IV feet overlap more than in 2A. In 2A, the l-word kaduda contained an IV-foot kadu and a statistical foot duda. In 2B, the l-word gukadu contains an IV-foot of kadu and a statistical foot of kadu. This simply means that the kadu syllable pair should be heard quite strongly. Indeed we can see in Table 10 that gukadu is the most selected word of any. Also, in Table 12, for trisyllabic-trisyllabic pairings, we see that gukadu again is the most selected word.

At the same time, kaduda is not selected nearly as highly as gukadu, which we cannot explain with either IV feet or statistical feet. Both types of feet only provide
information for *kadu*, which is contained in both. There is a danger that we are explaining noise, but there also could be a legitimate pattern here: As seen in Table 12, the three most selected words in the trisyllabic-disyllabic pairings are *gukadu*, *gakitu*, and *pigaki*. IV and statistical feet both provide reasons that those words would be among the most selected, but no reason they should be the top three specifically. There is, however, a commonality between all three words. We have focused upon SylProbs throughout this paper, but we have evidence that people can calculate probabilities for consonants as well (Bonatti et al., 2005; Newport & Aslin, 2004). By accident of language design, it happens that the consonant [g] is followed 100% of the time by [k]. If a [g] is heard, a [k] will always be the immediately next consonant. No similar pattern occurs within the A language. It is possible that participants are also noticing this pattern and thus are favoring words that include the *g-k* pair.

In sum, the combination of IV-feet and statistical feet explain the results of 2B in about the same way as 2A. Statistical feet support the disyllabic words and reinforce some of the IV-feet already heard, specifically *tipa* and *kadu*. One difference between 2A and 2B is that statistical feet (plus segmental probability for the consonants) can account for most of the results in Experiment 2B. In 2A, we needed IV-feet to understand why words like *gukadu* were selected. The IV-foot from the l-word *kaduda* in language A is *kadu*, which the boundary word *gukadu* contains. The statistical foot only reveals the *duda* part of *kaduda*, and then does not provide a reason for the erroneous selection of *gukadu*. In 2B, however, *kadu* is the IV-foot and the statistical foot.

This brings up the possibility of dropping IV feet from our explanation and relying entirely on syllabic and segmental probability. The results argue against this. As
just stated, Experiment 2A cannot be explained by statistical feet alone. The only way to get the participant to hear *kadu*, and hence *gukadu*, in 2A is to allow them to hear the entire word using SylProbs and not just the last two syllables. If we allow that, however, then performance should be higher and consistent, since the participant is hearing the entire word. Yet in both experiments, they continue to be confused about which trisyllabic words are the correct l-words. This suggests they are still only hearing pairs of syllables, not entire trisyllabic l-words.

The rhythmic beat in 2B plays largely the same role as in 2A. Its isochronous nature will draw rhythmic groupings, but the groupings will never fall upon a consistent structure. However, any binary grouping will always divide trisyllabic l-words into only one of two pairs (for instance, *gukadu* will always be *guka* or *kadu*). This repetition of only two syllable combinations within trisyllabic words, combined with the most stable IV-feet being inside trisyllabic words, will contribute to the trisyllabic bias we are seeing in Table 12.
6.1. Methodology

Much of the analysis of Experiment 2A and 2B is dependent on the notion that distributional patterns are cueing a prosodic boundary at the end of an l-word, and so the participant is segmenting the last two syllables of the l-word as a foot, and not the entire l-word. The participants in both experiments achieved a mean score of 0.55. One point of confusion in Experiment 2B was that the statistical feet occurred in both the trisyllabic l-words and the trisyllabic boundary words. For instance, the statistical foot *gaki* is contained within both the l-word *pigaki* and the boundary word *gakitu*. Due to this, two of the top three selected words were trisyllabic boundary words. If this analysis is correct, it can be tested. In Experiment 2D, the same training materials as in 2B were used. During testing, however, trisyllabic words that did *not* include statistical feet were employed as boundary word foils. We should see the proportion correct rise.

6.1.1. Participants

Eighteen adults participated in Experiment 2D. None had participated in the previous Experiments 1 or 2.

6.1.2. Stimuli

The identical training file from Experiment 2B was used for training in 2D. For the testing phase, the same disyllabic l-words as in 2B were employed. New trisyllabic boundary words were created, however. Whereas the trisyllabic boundary words for 2B contained the final 2 syllables of an l-word, the trisyllabic boundary words for 2D
contained the first two syllables of an l-word.\textsuperscript{21} Therefore, Experiment 2D boundary words no longer contain the statistical feet that we are positing participants perceive.

\textbf{6.1.3. Procedure}

The testing procedure was identical to Experiments 2A and 2B.

\textbf{6.2. Results – Analysis of Means}

Eighteen adults took part in Experiment 2D. None were excluded. As in all experiments, a chance response would have a mean score of 18 or a proportion correct of 0.5. Experiment 2D had a mean above 0.5 ($M=0.6142$, $SD=0.1174$). A plot of the density function appeared normal (Figure 19) and a Shapiro-Wilk Test for Normality confirmed this ($W=0.9694$, $p=0.7863$). Using a one-sample t-test the mean proportion correct for 2D was highly significant ($t(17)=4.1261$, $p=0.0004$).

![Density Curve](image)

Figure 9. Density curve for Experiment 2D.

\textsuperscript{21} Boundary words for Experiment 2D were: dututi, paguka, buda, dipiga, kubi, and kiba.
The prediction when designing Experiment 2D was that it would have a higher learning rate than Experiment 2B due to removing trisyllabic boundary words that contained statistical feet and which participants incorrectly selected. The mean proportion correct for 2D (\(M=0.6142\)) was higher than 2B (\(M=0.5523\)). A one-way ANOVA suggests the means are different with a p-value approaching significance (\(F(1,33)=3.587, p=0.0670\)).

6.3. Results – Analysis of Selections

In the interest of brevity, we will not perform the in-depth analysis of selections as in 2A and 2B. The overall results are the same in that IV feet are selected frequently, there is a bias for trisyllabic words, and disyllabic l-words are preferred over disyllabic boundary words. These results can be seen in Tables 14 and 15.

Table 14. Experiment 2D Selected Words

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>gukadu</td>
<td>0.127</td>
</tr>
<tr>
<td>tutipa</td>
<td>0.117</td>
</tr>
<tr>
<td>pigaki</td>
<td>0.105</td>
</tr>
<tr>
<td>daku</td>
<td>0.105</td>
</tr>
<tr>
<td>dututi</td>
<td>0.085</td>
</tr>
<tr>
<td>bibu</td>
<td>0.083</td>
</tr>
<tr>
<td>paguka</td>
<td>0.083</td>
</tr>
<tr>
<td>badi</td>
<td>0.077</td>
</tr>
<tr>
<td>buda</td>
<td>0.069</td>
</tr>
<tr>
<td>dipiga</td>
<td>0.060</td>
</tr>
<tr>
<td>kubi</td>
<td>0.052</td>
</tr>
<tr>
<td>kiba</td>
<td>0.035</td>
</tr>
</tbody>
</table>

Even though participants in 2B and 2D went through an identical training, we see something new in the trisyllabic-trisyllabic test items. In 2B, participants did not show an
ability to discriminate the trisyllabic l-words from trisyllabic boundary words. Of the three most selected trisyllabic words in 2B, two of them were boundary words. In 2D, however, the three trisyllabic l-words are chosen over the three trisyllabic boundary words, though the l-word pigaki is actually chosen at the exact same rate as the boundary word paguka. The statistical feet are not contained in the boundary words for 2D, so it appears in the test that the participants know the entire l-word. After all, in the trisyllabic-trisyllabic pairings, they prefer all three l-words. However, they might still only know two syllables of each l-word. The lack of knowledge is not revealed by the test words of 2D, but the lack of knowledge was revealed in 2B.

Table 15. Experiment 2D Selected Words Grouped by Word Pairings

<table>
<thead>
<tr>
<th>Disyllabic-Disyllabic</th>
<th>Trisyllabic-Disyllabic</th>
<th>Trisyllabic-Trisyllabic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>Frequency</td>
<td>Word</td>
</tr>
<tr>
<td>daku (L)</td>
<td>0.222</td>
<td>gukadu (L)</td>
</tr>
<tr>
<td>badi (L)</td>
<td>0.203</td>
<td>pigaki (L)</td>
</tr>
<tr>
<td>bibu(L)</td>
<td>0.190</td>
<td>tutipa (L)</td>
</tr>
<tr>
<td>buda (B)</td>
<td>0.163</td>
<td>dututi (B)</td>
</tr>
<tr>
<td>kubi (B)</td>
<td>0.131</td>
<td>paguka (B)</td>
</tr>
<tr>
<td>kiba (B)</td>
<td>0.092</td>
<td>daku (L)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dipiga (B)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bibu (L)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>badi (L)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>buda (B)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>kubi (B)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>kiba (B)</td>
</tr>
</tbody>
</table>

Next we turn to Experiment 3. Experiments 3 uses the same language and testing foils as Experiment 2. However, during training the trisyllabic l-words are shortened so that they have the same duration as disyllabic l-words.
CHAPTER 7: EXPERIMENT 3 – RESTORING ISOCHRONOUS WORD ONSETS

7.1. Methodology

Experiment 2 had uniform syllable durations but varying word durations due to the use of both di- and trisyllabic words in the artificial language. Experiment 3 uses the same languages and procedures as Experiments 2A and 2B; however, the durations of vowels have been modified so that word durations are now uniform, but syllable durations vary.

7.1.1. Participants.

Thirty-eight participants took part in the experiment; eighteen in language A and nineteen in language B. No participants in Experiment 3 participated in any other experiment.

7.1.2. Stimuli.

Languages A and B were re-used from Experiment 2. However, the vowels of trisyllabic words were shorter than the vowels of disyllabic words. Syllable duration in disyllabic words remained 330 ms. Syllable duration was shortened in trisyllabic words to be 220 ms. This results in words that are all 660 milliseconds in length. The same number of word tokens occurred in Experiments 1, 2, and 3. Experiment 3’s training block was 7 minutes in length as in Experiment 1.

Due to the differences in syllable duration for the training materials in Experiment 3, a testing issue arises that was absent in Experiments 1 and 2. Namely, what should the durations of test syllables be? If the testing materials used either 330 ms or 220 ms (the durations in the training materials), the participants may select the test words based
entirely on the speech rate, not the syllabic content. As a solution, the average of the vowel durations in training (190 ms and 300 ms) was used for the test word vowel duration so that participants would not choose disyllabic or trisyllabic words simply based upon a match to the durations within the training speech. This yielded a 275 ms test syllable. Both disyllabic and trisyllabic words had this same duration.

7.1.3. Procedures

Experiment 3 followed the same procedures as Experiments 1 and 2.

7.2. Results – Analysis of Means

Eighteen people took part in Experiment 3A and nineteen in Experiment 3B. During analysis, one participant’s results were excluded from 3B due to being more than 2 standard deviations from the mean. Therefore, n=18 for both 3A and 3B. Participants in Experiment 3A ($M=0.5370$, $SD=0.0707$) scored slightly lower than those in 2A ($M=0.5529$, $SD=0.0528$). Those in 3B, however, did better than in any experiment so far ($M=0.6543$, $SD=0.1258$). Before statistical analysis then, it appears that behavior in 3A was very different than in 3B. This does not follow the pattern of Experiments 1 and 2, in which we saw no significant difference between languages. This is surprising since Experiment 3 uses the same training language as Experiment 2.

The next step was to assess the normality of the distributions. Plots of each experiment’s density curves are found in Figure 10. Experiment 3A’s density function appears roughly normal with a slight skew. Experiment 3B, however, potentially reveals a bimodal distribution. Shapiro-Wilk tests for normality confirm these impressions. Experiment 3A remains normal ($W=0.9211$, $p=0.1350$), while 3B is not ($W=0.887$, $p=0.0342$).
A histogram of 3B provides further confirmation of a bimodal distribution (Figure 11, right panel). There appears to be a clean break around a proportion of 0.64 with all
values occurring either to the left or to the right of this break point. The participants were therefore split into 3B-High and 3B-Low sets. All who achieved a score greater than 0.64 were placed in the High group and all scoring below 0.64 were placed in the Low group. This resulted in 7 participants in the High group and 11 participants in the Low group with very disparate means but similar variance (High: $M=0.7976$, $SD=0.04454$; Low: $M=0.5613$, $SD=0.0466$).

![Density Curves](image1)

**Figure 12.** Density curves for Experiments 3B-High and 3B-Low.

We can now assess the normality and variance of each distribution separately. The density functions for 3B-High and 3B-Low are found in Figure 12. Their appearance is quite normal at this point and Shapiro-Wilk Tests for Normality confirm this impression for both 3B-High ($W=0.9671$, $p=0.8766$) and 3B-Low ($W=0.945$, $p=0.5805$). We have no evidence that the High and Low groups do not follow a normal distribution.
We next perform a parametric variance test as well ($F(6,10)=0.9124$, $p=0.9528$), which indicates the two distributions do not have different variances. Assumptions are met to run a one-way ANOVA to compare means, which indicates that the means are very unlikely to be the same ($F=(1,16)=111.84$, $p<.0001$). This is not a surprising result since the High and Low data sets have no points of overlap.

We can now assess learning in both High and Low groups. A one-sample t-test confirms that there was significant learning in the High group ($t(6)=17.67$, $p<.0001$) and in the Low group ($t(10)=4.49$, $p=0.0011$). Naturally, the High group’s learning is far higher than the Low group’s, a 24% difference. Before we try to understand why Experiment 3B has such a bimodal distribution when there was little evidence of it in the first two experiments, it is useful to compare 3B further to the other experiments.

In Experiments 1 and 2, we saw no significant difference between the A and B languages. Since 3A and 3B use the same languages as 2A and 2B, which had identical means, it is possible that 3A will also be comparable to 3B-Low, the group of 11 participants who did not perform exceptionally though they did perform above chance. We already know that both 3A and 3B-Low follow normal distributions. A variance test also confirms that they have similar variance ($F(17,10)=2.296$, $p=0.1829$). A one-way ANOVA then confirms that there is no significant indication that the means of 3A and 3B-Low are different ($N=18+11=29$; 3A: $M=0.5370$; 3B-Low: $M=0.5613$; $F(1,27)=1.1773$, $p=0.2875$). We then collapse 3A and 3B-Low together, as we did with 1A and 1B and with 2A and 2B before. We test this combined 3A/3B-Low group for normality with Shapiro-Wilk ($W=0.9418$, $p=0.1116$), and then a one-sample t-test confirms learning ($M=0.5469$, $SD=0.0630$. $t(28)=4.0094$, $p=.0004$).
Earlier we saw that Experiment 2 showed a lower degree of learning than Experiment 1, but that the difference did not reach the <0.05 threshold. How does Experiment 3, as represented by the 3A/3B-Low group, compare with Experiments 1 and 2? At this point, 3 pair-wise comparisons are called for. Multiple comparisons of this sort increase the likelihood of a Type I error. Two common adjustments for this are the Bonferroni correction and Tukey’s Honestly Significant Difference. Tukey’s HSD generally has more power to discern differences than the Bonferroni correction, but it requires similar N in each experiment. However, our experiments do not have similar participant numbers across experiments (Exp. 1: N=24; Exp. 2: N=34; Exp. 3: N=29). Therefore, we will use the Bonferroni correction, which means that we should not accept a difference as significant unless it is less than 0.05/3=0.0167. A one-way ANOVA provides no evidence that Experiment 2 and Experiment 3 (3A/3B-Low) have different means ($F(1,61)=0.1191, p=0.7312$). A one-way ANOVA provides the same result for Experiment 1 and Experiment 3 ($F(1,51)=2.3884, p=0.1284$).

In sum, what we have so far is that, if the high group from 3B is excluded, all languages demonstrate learning and none is significantly different than the others, despite the fact that the means for both Experiment 2 and Experiment 3 are lower than the mean for Experiment 1. Figure 13 summarizes these findings.

The fact that only 3B-High is significantly different from other groups suggests that its participants are segmenting the language in a way that is different from the other groups. This is surprising if only transitional probabilities are available as a cue, as the distributional pattern of syllables is identical in every experiment in all languages. In Section 7.3, we turn to a detailed analysis of the pattern of word selections to try to assess
precisely how the groups earn the scores they do. We will see, for instance, that while the mean proportion correct for 2A and 3A are very close, the words they select are quite different – despite the fact that the same l-words and boundary words occur in both tests.

Figure 13. Mean proportion correct for Experiments 1, 2, and 3.

7.3. Results – Analysis of Selections

7.3.1. Analysis of 3A

The principles guiding our interpretation of Experiment 3 are the same as they were for Experiment 2: The listener will segment the speech stream immediately and continuously and will always apply their native language’s phonology. This means that
the same concepts of IV-feet, statistical feet, and rhythmic grouping will be relevant. As a reminder, we will state the primary IV-feet, those feet that repeat within words, in (41).

(41) IV-Feet: tipa, pabi, kadu, kitu, kuba

The statistical feet will not emerge until some time passes while listening to the training language, so we will wait to present those.

One new thing is that the rhythmic groupings will be different in Experiment 3 than in Experiment 2. Experiment 3 contains even word durations, and the participant’s internal oscillators could become entrained with this pattern. Example (42) provides a sample of language A with a bar before each word beat to evoke the idea of meter.

(42) | tipabi | kuba | gakitu | bupi | digu | kaduda

The rhythmic groupings in this case, unlike in Experiment 2, mark out all l-words completely. If participants can hear this rhythm, surely all participants should learn the language perfectly – but they do not. Why?

The training language of Experiment 3 shortened the durations of syllables in trisyllabic l-words so that they are the same duration as disyllabic l-words. This manipulation accomplishes that goal. However, having different syllable durations for the two types of word creates other possible cues for the meter of the speech as well. As discussed in Section 3.3.1, when a trisyllabic l-word and disyllabic l-word are adjacent, the syllables at the adjoining edge are different in length, which can have an impact on perception. If a trisyllabic l-word occurs first, followed by a disyllabic l-word, then it could sound like a syllable has been lengthened. Word final lengthening is a known marker for word segmentation and can mark the end of a prosodic boundary. Similarly, if a disyllabic l-word word occurs first and is followed by a trisyllabic l-word, which has a
shorter syllable, then that could be perceived as a stressed syllable followed by an unstressed syllable. Duration is one acoustic cue for stress in American English.

Examples (24) and (25) from Section 3.3.1 are copied below as (43) and (44)

(43) kadudadi:gu:
(44) di:gu:kaduda

Pairs such as (43) and (44) will occur sporadically through the training language due to the random concatenation of words. Since a word cannot recur after itself, a trisyllabic l-word is more likely to be followed by a disyllabic l-word than another trisyllabic (60%/40%), and vice versa. It is possible that pairs such as (43) will be perceived as containing final lengthening on the di syllable, since di is longer than the syllable that precedes it. Assuming binary feet, dadi: would be the parsed foot. Example (44) in turn could be perceived as containing a stress cue on the syllable gu, since it is longer in duration than the next syllable. This would cue the binary foot gu:ka.

Pursuing lengthening first, we can construct a phonological parse of (43). First, perceived lengthening on the syllable di: yields dadi: as a foot. That leaves three syllables unparsed: kadu before the dadi: foot and a spare syllable gu afterwards. It seems very likely that kadu would end up as another foot. On the most basic level, ka and du are simply the two syllables before the parsed dadi: and so can be grouped. Moreover, there are two other cues that ka, of kadu, is the beginning of a foot. The syllable ka begins with an aspirated voiceless stop and is an IV-foot. Additionally, there is a regular beat occurring on the beginning of each trisyllabic l-word, which could also mark the start of a metrical foot.
A complete metrical parse of the first 4 syllables is then possible. Example (45) presents the language sample in bold, the cues for feet above the words, and the resulting metrical feet below the words (see Chapter 2 for a review of metrical phonology).

Voice: -v +v +v +v +v
Beat: x x

(45) ka du da di: gu:
| x | x |
| x x | x x |
Foot# 1 2

A beat and aspiration occur on ka. These cues both mark the beginning of foot #1. The perceived lengthening on di then marks the end of foot #2. A slightly different parse is also available when we consider that another beat occurs on di. Since a beat occurs there, the participant might find the same feet but consider foot #2 to be iambic. This is presented in (46). The only difference between (45) and (46) is that the stress moved from the first syllable to second syllable of foot #2.

Voice: -v +v +v +v +v
Beat: x x

(46) ka du da dii gu
| x | x |
| x x | x x |
Foot# 1 2

It is somewhat difficult to know what would happen with the gu syllable at the end of (45) and (46). One possibility is that it would remain unparsed or extrametrical. Alternatively, since it does have a longer duration than a syllable in a trisyllabic l-word, it might be placed into a foot of its own. The second alternative is easier to see if we string multiple di- and tri-syllabic l-words together (47).

(47) ka du da di: gu: ti pa bi ku: ba: |
| x | x | x | x | x | x |
| x x | x x | x x | x x | x x | x x |
Foot# 1 2 3 4 5 6
A drawback to the iambic parse is that its stress could now clash with the monosyllabic foot that follows. If so, the stress would move back to the first syllable: *dádi*, instead of *dadi*. In short, a fairly regular metrical structure can be found in the training language for 3A using perceived lengthening, the rhythm, and IV-feet.

If such a structure is heard, two primary sets of metrical feet will be perceived. One set is composed of the last syllable of the trisyllabic l-word and the first syllable of a disyllabic l-word. This corresponds to foot #2 in (46) and (47). The possible words that would be heard this way are: *dadi, daku, dabu, bidi, biku, bibu, tudi, tuku*, and *tubu*. Since these feet are always across a word boundary, they are not heard as frequently as those within a word boundary. The word-internal feet that come from the parsing in (47) are just the first two syllables of each trisyllabic l-word, corresponding to foot #1 in (47). The l-words are *kaduda, tipabi*, and *gakitu*, so the word-internal feet are *kadu, tipa*, and *gaki*.

The parsing of (47), what we might call the final lengthening parse, is not quite as stable as it might seem, however. If disyllabic-trisyllabic words always alternated, then one could parse the whole stream this way. They do not alternate, of course. The language was composed by concatenating words randomly, except that a word cannot follow itself. This means that there is only a 60% chance that a disyllabic l-word will follow a trisyllabic l-word. (48) presents a sample of language A in which 3 trisyllabic words are followed by two disyllabic ones.

(48)  kadudatipabikadudadi:gu:bu:pi:

There is only a single pair of syllables in which word final lengthening would be perceived in this stretch of speech, and so the lengthening notion would be of little help.
In such stretches, the participant would need to use the same attributes they used in 2A: IV-feet and rhythmic grouping.

Table 16. Experiment 3A Selected Words

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>kaduda</td>
<td>0.114</td>
</tr>
<tr>
<td>tipabi</td>
<td>0.113</td>
</tr>
<tr>
<td>tutipa</td>
<td>0.106</td>
</tr>
<tr>
<td>gukadu</td>
<td>0.093</td>
</tr>
<tr>
<td>bupi</td>
<td>0.088</td>
</tr>
<tr>
<td>gakitu</td>
<td>0.086</td>
</tr>
<tr>
<td>kuba</td>
<td>0.077</td>
</tr>
<tr>
<td>badi</td>
<td>0.074</td>
</tr>
<tr>
<td>pigaki</td>
<td>0.074</td>
</tr>
<tr>
<td>bibu</td>
<td>0.066</td>
</tr>
<tr>
<td>digu</td>
<td>0.059</td>
</tr>
<tr>
<td>daku</td>
<td>0.049</td>
</tr>
</tbody>
</table>

To sum, using lengthening, the participant should most often hear the three feet *kadu*, *tipa*, and *gaki*. Using IV-feet, they should hear *tipa*, *pabi*, *kadu*, and *kitu*.

Additionally, they will hear some feet that straddle a word boundary due to the perceived lengthening. And, finally, we will need to consider the statistical feet. As with 2A, this list will include the three disyllabic l-words (*kuba, digu, bupi*) plus the last two syllables of each trisyllabic l-word (*duda, pabi, and kitu*). We can now compare this to the words participants selected in Experiment 3A (Tables 16 and 17).

The lengthening notion supports hearing the feet *kadu*, *tipa*, and *gaki* with *tipa* and *kadu* also receiving support from being both on a beat and voiceless. This is the parse seen in (47). Indeed, 5 of the 6 words selected over chance include this set of feet. The only exception is that the disyllabic l-word *bupi* is also selected over chance (Table 16).
Moreover, the participants only show moderate ability to distinguish between trisyllabic words that both contain the same foot. If we look at the trisyllabic vs. trisyllabic test items (Table 17), the l-word tipabi is chosen at a rate of 0.198 and the boundary word tutipa is chosen at a rate of 0.191, a very small difference. This confusion between trisyllabic words is something we saw in 2A as well.

Table 17. Experiment 3A Selected Words Grouped by Word Pairings

<table>
<thead>
<tr>
<th>Disyllabic-Disyllabic Word</th>
<th>Frequency</th>
<th>Trisyllabic-Disyllabic Word</th>
<th>Frequency</th>
<th>Trisyllabic-Trisyllabic Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>bupi (L)</td>
<td>0.204</td>
<td>kaduda (L)</td>
<td>0.132</td>
<td>tipabi (L)</td>
<td>0.198</td>
</tr>
<tr>
<td>kuba (L)</td>
<td>0.185</td>
<td>tipabi (L)</td>
<td>0.127</td>
<td>kaduda (L)</td>
<td>0.191</td>
</tr>
<tr>
<td>badi (B)</td>
<td>0.167</td>
<td>tutipa (B)</td>
<td>0.117</td>
<td>tutipa (B)</td>
<td>0.191</td>
</tr>
<tr>
<td>bibu (B)</td>
<td>0.167</td>
<td>gukadu (B)</td>
<td>0.104</td>
<td>gukadu (B)</td>
<td>0.160</td>
</tr>
<tr>
<td>digu (L)</td>
<td>0.142</td>
<td>gakitu (L)</td>
<td>0.096</td>
<td>gakitu (L)</td>
<td>0.154</td>
</tr>
<tr>
<td>daku (B)</td>
<td>0.136</td>
<td>pigaki (B)</td>
<td>0.096</td>
<td>pigaki (B)</td>
<td>0.105</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bupi (L)</td>
<td>0.074</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>badi (B)</td>
<td>0.065</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>kuba (L)</td>
<td>0.062</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>bibu (B)</td>
<td>0.049</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>digu (L)</td>
<td>0.046</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>daku (B)</td>
<td>0.031</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Something we did not see, however, in 2A is confusion about the disyllabic l-words. In 2A, when a disyllabic word was paired with a disyllabic word, participants reliably chose all l-words more frequently. In 3A, however, one l-word, digu, is selected less often than two boundary words. This is predicted by the lengthening account of metric parsing, because the perception of lengthening starts by attaching the first syllables of a disyllabic l-word with the last syllable of a trisyllabic l-word. This effectively splits a disyllabic l-word making it difficult to recognize in testing. The lengthening parse only leaves the participant with one syllable with which to recognize the disyllabic word later.
The preference for the disyllable l-word *bupi* could be further evidence this is happening. If the first syllable *bu* is split from *bupi* and attached to the preceding trisyllabic word, this leaves just *pi* as a possible metric degenerate foot unto itself (again see (47)). The syllable *pi* starts with a voiceless stop, which could reinforce the notion of *pi* as a foot. Since all syllables are unique to their l-word, participants could be using just *pi* to correctly select *bupi*. In 2A, disyllabic l-words were not reliably split by the lengthening perception, and there we did not see *bupi* being selected more than all other disyllabic l-words. The other disyllabic l-words in 3A happen to have voiced consonants for their second syllables so they may not be as reliably heard.

If we switch our attention to the trisyllabic-disyllabic pairings (Table 17), we first notice the same sort of trisyllabic bias that we saw in Experiment 2. When a trisyllabic is tested against a disyllabic, the trisyllabic words are reliably selected. There could be several reasons for this. The most likely is that the lengthening parse revealed the first part of each trisyllabic word, and so those sound the most familiar to the participant. Another possible reason has to do with the actual duration of the words of the test. As the reader will recall from Section 5.3.2, a brief experiment provided only the testing materials to 5 participants to see if they preferred certain words. While there was no significant bias for the words of a specific language, there was a strong preference for trisyllabic words when the duration of a syllable was 245 ms, instead of the 330 ms used in Experiment 2. Could the participants be choosing trisyllabic words on that basis alone?

While it is possible, the data suggest not. First, as we will soon see, when a language is truly learned, as it was by the 3B-High group, the trisyllabic bias largely vanishes. If the bias for trisyllabic words is at play here, it is likely only relevant because
the participants are already having trouble distinguishing l-words and boundary words. It is not obscuring the underlying knowledge. This is confirmed by another experiment, not reported here, in which participants were presented with the same training and testing items of Experiment 3A, except that the syllables in testing were 330 ms in duration, instead of 245 ms. Participants in that experiment performed equivalently to those in 3A.

Looking through the word selections in Tables 16 and 17, the overall picture is one of participant confusion. The boundary words are continually mixed with the l-words. This matches the overall result in which learning was at 54%. While this score is significantly different from chance, it is hardly robust.

In the explanations of Experiment 2, however, we had a role for statistical feet, which have played little role in the results of Experiment 3A. This seems to be the case. We have no strong evidence in 3A that statistical feet were important to the participants’ word selections. The most likely explanation is that the lengthening, aspiration, and rhythmic cues overwhelmed the statistical cue. This matches previous research in which prosodic segmentation cues were pitted against statistical ones. In those cases, participants chose the segmentation compatible with the prosodic parse, not the statistical one (Johnson & Juszczyk, 2001; Shukla, Nespor, & Mehler, 2007). This could be due to participants giving cues such as lengthening more weight than distributional ones, or it could be due to the fact that lengthening is immediately available to use in segmentation, while the statistical structure can only unfold over time.

\[\text{Wald’s } Z_{\text{first syllable of l-word}} = 2.615, p=0.0089; \text{Wald’s } Z_{\text{second syllable of l-word}} = 3.0101, p=0.0026; \text{Wald’s } Z_{\text{first syllable of boundary word}} = -3.106, p=0.0019; \text{Wald’s } Z_{\text{second syllable of boundary word}} = -2.008, p=0.0446.\]

\[\text{Due to the fit of the lengthening model with the data, we have not concentrated as much on IV-feet as we did in Experiments 2A and 2B. However, they remain significant according to a mixed effects model: First syllable of an l-word (Wald’s } Z_{\text{first syllable of l-word}} = 2.615, p=0.0089; \text{Second syllable of an l-word (Wald’s } Z_{\text{second syllable of l-word}} = 3.0101, p=0.0026; \text{First syllable of a boundary word (Wald’s } Z_{\text{first syllable of boundary word}} = -3.106, p=0.0019; \text{Second syllable of a boundary word (Wald’s } Z_{\text{second syllable of boundary word}} = -2.008, p=0.0446).}\]
Another possible reason that statistical feet do not play a large role relates to the Information-Duration Hypothesis, which argued that rises in uncertainty would be experienced as a rise in duration. If this were the case, the unpredictability that occurs at a word boundary would be experienced as a slight lengthening on the last syllable before the boundary. Recall that Experiment 3A is the first time that participants have not segmented disyllabic words successfully (Table 17). When IDH lengthening happens with disyllabic l-words, the result will be a perceived lengthening on the second syllable. This could actually support the overall lengthening parse of (47). The second syllable of a disyllabic word is the leftover syllable occurring right after the lengthening (Feet #3 and #6 in (47)). One way to parse it would be as a monosyllabic degenerate foot (Section 2.1.2). The longer the syllable sounds, the more likely the monosyllabic foot becomes. In this case then, knowing the statistical structure could actually support the lengthening based parse.

When we move to trisyllabic l-words, the perceived extra duration from statistical uncertainty would occur on the final syllable of a trisyllabic l-word. According to the metrical parse of (47), however, this is also the location of perceived stress, since the next syllable is heard as final lengthening. Increased perceived duration on that syllable would again reinforce the lengthening parse.

We have one more loose end to tie up before moving to the final experiment – Experiment 3B. Earlier, it was mentioned that the differences in syllable duration between disyllabic and trisyllabic l-words could trigger a perception of stress. This could

23 Recall from (47) that kadudadi:gu: is parsed by lengthening as kadu|dadi::|gu:. Assuming trochaic feet, this places stress on da of dadi:.
happen when the disyllabic is followed by a trisyllabic. Table 18 lists all the possible
combinations. The adjacent syllables between the two l-words are highlighted.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>digu</td>
<td>kaduda</td>
</tr>
<tr>
<td>digu</td>
<td>tipabi</td>
</tr>
<tr>
<td>digu</td>
<td>gakitu</td>
</tr>
<tr>
<td>kuba</td>
<td>kaduda</td>
</tr>
<tr>
<td>kuba</td>
<td>tipabi</td>
</tr>
<tr>
<td>kuba</td>
<td>gakitu</td>
</tr>
<tr>
<td>bupi</td>
<td>kaduda</td>
</tr>
<tr>
<td>bupi</td>
<td>tipabi</td>
</tr>
<tr>
<td>bupi</td>
<td>gakitu</td>
</tr>
</tbody>
</table>

The resulting feet are: guka, guti, guga, baka, bati, baga, pika, piti, piga. Two of these
combinations occur in boundary words for Experiment 3A, *guka* in *gukadu* and *piga* in
*pigaki*. We would expect participants to false alarm on those two boundary words if they
are hearing the syllables within using this possible stress cue. The boundary word *pigaki*,
containing *piga*, was chosen the least of all tri-syllabic words. The boundary word
*gukadu*, containing *guka*, was selected fairly robustly, but this can be explained by the
presence of the sequence *kadu* within *gukadu*. Moreover, if we use *guka* to explain the
selection of *gukadu*, then we have no way to explain the even more common selection of
*kaduda*, because *guka* is formed by taking the *ka* syllable away from *kadu*. The only
viable alternative would be that both *guka* and *kadu* are heard, *guka* from the stress
perception and *kadu* as an IV-foot. However, since we see no evidence of a preference
for the other possible false alarm, *pigaki*, the simpler explanation is that stress perception
is not triggering false alarms.

Why do we not have evidence that this sort of stress cue is used in segmentation?
First, this matches Saffran et al.’s (1996) results that showed lengthening was a better
segmentation cue than stress. Moreover, note that a rhythmic beat occurs right after the
possible stressed syllables in Table 18. This could produce stress clash. Taking the *digu-kaduda* example from Table 18, if a beat occurs on *ka* (the start of a trisyllabic l-word) and this is evidence of a metrical foot, then another stress cannot occur on the immediately previous syllable. (49) displays the stress clash. The syllable *gu:* is marked as stressed due to the longer duration, but *ka* is marked as stressed due to the beat on the word onset.

\[
\begin{array}{cccc}
\text{di:} & \text{gu:} & \text{ka} & \text{du} & \text{da} \\
| & x & x & | & x & x & x \\
\end{array}
\]

The only way to avoid this stress clash would be for *gu,* of *digu,* to be a monosyllabic foot by itself, which is what was proposed in (47), the so-called lengthening parse, not the stress parse.

In summary, a perceived final lengthening due to a long syllable from a disyllabic l-word following the relatively shorter syllable of a trisyllabic l-word, combined with a beat on each word onset and the perception of aspiration, can create a relatively stable metrical structure for language 3A. Statistical calculations are either not strong enough to overcome this perceived structure or happen to reinforce it. The result is only moderate learning based upon the perceived lengthening parse. This “moderate” learning would arise from fairly subtle preferences, such as preferring the test word that starts with a perceived foot rather than one that simply contains it; i.e., if *kadu* was perceived, participants choose *kaduda* more than *gukadu.* Moreover, even though the mean proportion correct in 3A is quite similar to 2A, when we look at the precise selections, we can see a rather different perception of the speech stream.
7.3.2. Analysis of Experiment 3B-Low

The important question with 3B is why the 3B-High group emerged here and nowhere else. To understand that, we will apply the same sort of analysis on 3B-Low that we have on 2A, 2B, 2D, and 3A. The model of IV-feet, statistical feet, rhythm, and lengthening should explain that result as well. We can then show how 3B-High is different. We start by listing the l-words (50) and IV-feet (51), which are the same as in 2B.

(50) tutipa gukadu pigaki badi bibu daku
(51) titu tipa kadu piga

The IV-feet listed in (51) are only those that occur word internally. Others will occur across a word boundary, but be heard less frequently. Experiment 2B also saw a possible role for transitional probability between consonants. In particular, the chance of a [k] following a [g] is 100% within Language B.

One of the major differences between Experiments 2B and 3B is the possible perception of final lengthening on the first syllable of a disyllabic l-word that follows a trisyllabic word, as just documented in 3A. Example (52) demonstrates how this would work for two words from language B, gukadu followed by daku.

Voice: +v -v +v +v -v
Beat: x x
(52) gu ka du da: ku:
| x | x |
| x x | x x |
Foot# 1 2

As with Experiment 3A, the relatively longer syllable on da would be perceived as the end of a foot, in this case, duda:. The two syllables before it would also become a foot,
guka. The syllable gu is also marked as a possible beginning of a foot because the beat falls there. Example (53) applies the same principles to a stretch of 4 words.

(53) gu ka du da: ku: tu ti pa bi: bu: |
    | x   | x   | x   | x   | x   | x   |
    | x   | x   | x   | x   | x   | x   |
Foot# 1 2 3 4 5 6

Several feet emerge from (53). Two feet are marked by the syllable perceived as lengthened (feet such as #2 and #5, duda: and pabi:), two more feet from being marked with a beat and preceding the “lengthened” feet (feet such as #1 and #4, guka and tuti), and two more from the leftover last syllable of a disyllabic l-word (feet such as #3 and #6, ku and bu). In Experiment 3A, it was the syllable pair inside a trisyllabic word (from syllables 1 and 2) that should be heard the most often. These word-internal feet do not cross a word boundary and they are marked by both a beat and the lengthening on the next foot. Moreover, in 3A, two of those feet were tipa and kadu, which are marked by aspiration on a voiceless stop.

The support for the word-internal feet is not quite as strong in 3B. (1) In the first l-word of (53), gukadu, the first voiceless stop happens to be on the second syllable, not the first. (2) A related issue occurs on another trisyllabic l-word, tutipa. The segment [t] happens to be highly aspirated in the output of the MBROLA synthesizer, making it particularly salient. The l-word tutipa contains a [t] on both the first and second syllable. This suggests that some participants could be hearing tuti as an IV foot and tipa as another IV-foot. In other words, an aspirated, highly salient [t] could be splitting apart the tuti foot. (3) Finally, our third trisyllabic l-word, pigaki, does indeed have a voiceless stop on the first syllable and not the second. This would lend support to hearing the first two syllables, pi and ga, as a foot in the lengthening context. However, we also recall that k
follows \( g \) with probability 1, which may influence some listeners to cluster \( ga \) and \( ki \) in 
\( pigaki \), instead of splitting \( ga \) away to become part of \( piga \). None of this is to argue that 
the lengthening parse will not be a possible segmentation of 3B as it was for 3A. 
However, in all three trisyllabic l-words, there are reasons a participant might wish to 
group the second two syllables instead of the first two. Almost all cues reinforced 
grouping the first two syllables in 3A.

Let us compare this output of IV-feet, perceived lengthening, and a rhythmic beat 
to the words that participants selected in 3B. As the reader will recall, 3B revealed a 
bimodal distribution: a Low group whose mean score was 56% and a High group whose 
mean score was 80%. We will first look at the 3B-Low group to see if it follows the 
patterns we have seen in experiments so far, and then compare this to 3B-High.

Table 19. Experiment 3B-Low Selected Words

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>gukadu</td>
<td>0.106</td>
</tr>
<tr>
<td>tutipa</td>
<td>0.104</td>
</tr>
<tr>
<td>pigaki</td>
<td>0.101</td>
</tr>
<tr>
<td>bibu</td>
<td>0.091</td>
</tr>
<tr>
<td>kuba</td>
<td>0.091</td>
</tr>
<tr>
<td>badi</td>
<td>0.088</td>
</tr>
<tr>
<td>kaduda</td>
<td>0.086</td>
</tr>
<tr>
<td>tipabi</td>
<td>0.076</td>
</tr>
<tr>
<td>daku</td>
<td>0.073</td>
</tr>
<tr>
<td>bupi</td>
<td>0.071</td>
</tr>
<tr>
<td>gakitu</td>
<td>0.071</td>
</tr>
<tr>
<td>digu</td>
<td>0.043</td>
</tr>
</tbody>
</table>

Using the lengthening parse, the participant will hear the first two syllables of 
each trisyllabic l-word (\( guka \), \( tuti \), and \( piga \)) as a foot. The three most selected words
overall for the 3B-Low group contain precisely those syllables (Table 19). In 3A, this type of foot was contained within boundary words, but in 3B that is not the case. As a result, the other trisyllabic boundary words are not selected nearly as often. Indeed, only one is selected (barely) above chance. The boundary word *kaduda* is selected at a rate of .086, while chance is 0.83. All others are below that. The preference for trisyllabic l-words over trisyllabic boundary words is also seen when we look at the trisyllabic-trisyllabic pairs in Table 20. For those test items, the l-words are reliably selected more than boundary words. Note that this is a very different pattern than seen in 2B. 2B uses the same language as 3B, but in 2B participants had a hard time telling the trisyllabic words apart. Importantly, this does not mean that the participants in the 3B-Low group must have parsed the entire trisyllabic word, while those in 2B did not. The lengthening parse allows you to grab the first two syllables of a trisyllabic word, and that is sufficient to discriminate all trisyllabic l-words and boundary words. In 2B, perceived lengthening was not a factor and so they were selecting the latter two syllables in a trisyllabic word, which are in the boundary words, based on statistical and IV-feet.

<table>
<thead>
<tr>
<th>Table 20. Experiment 3B-Low Selected Words Grouped by Word Pairings</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Disyllabic-Disyllabic</strong></td>
</tr>
<tr>
<td>Word</td>
</tr>
<tr>
<td>bibu (L)</td>
</tr>
<tr>
<td>badi (L)</td>
</tr>
<tr>
<td>kuba (B)</td>
</tr>
<tr>
<td>daku (L)</td>
</tr>
<tr>
<td>bupi (B)</td>
</tr>
<tr>
<td>digu (B)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
We have some evidence that the participants in 3B-Low do not have a complete parse of the whole trisyllabic l-word in the trisyllabic-disyllabic pairs (Table 20). In these trisyllabic-disyllabic pairs, the participants are judging whether a trisyllabic word sounds more like a word of the language than a disyllabic word, and vice versa. In this context, participants do choose *kaduda*, a boundary word in 3B, above chance. This could be because they have heard the *kadu* IV-foot inside the l-word *gukadu*. However, *kaduda* also contains *duda*, which is one of the feet that comes the lengthening parse that crosses word boundaries: *gukadu daku* is parsed as *guka|duda|ku*. So the lengthening parse correctly predicts selections such as *kaduda*.

But why not select *kaduda* as often in the trisyllabic-trisyllabic pairs as well? One possible reason is relative strength. The *guka* foot has a chance to be heard every single time the word *gukadu* appears, while the *duda* foot will only be heard when *gukadu* is followed by *daku*. This will only happen in 20% of the appearances of *gukadu*. (33% if only following disyllabic l-words are considered.) Similarly the lengthening parse will actively split disyllabic l-words apart, so that they sound even less like something heard during training. Thus, in trisyllabic-trisyllabic pairs, the participants can distinguish *guka* from *duda*, but in trisyllabic-disyllabic pairs, *duda* sounds more likely than *daku*.

Finally, we see some confusion about which disyllabic words are the correct ones when we look at the disyllabic-disyllabic pairs in Table 20. The *kuba* boundary word is selected as often as any other l-word except one. The most exceptional pattern in the disyllabic-disyllabic pairings is the very active avoidance of the boundary word *digu*. It is chosen far less than any other word. One possible reason for this is that the boundary word *digu* will occur when the l-word *badi* is followed by *gukadu*. We have seen that
gukadu is the most selected word of the group (Table 19). Perhaps guka is such an obvious foot to participants that the last syllable of badi is rarely heard as grouped with the first syllable of guka. Note that this would be a possible “stress foot” since it is a disyllabic l-word followed by a trisyllabic l-word. The avoidance of digu strongly suggests that this possible stress cue is not being used to parse.

The results for the 3B-Low group then are generally interpretable in the same terms as 3A. The results are fairly well explained by perceived lengthening when disyllabic l-words follow trisyllabic l-words, plus IV-feet and segmental probability. We now turn to the 3B-High group to complete the primary analysis of Experiments 2 and 3.

7.3.2. Analysis of Experiment 3B-High

Tables 21 and 22 provide the words selected for the 3B-High group. Quite simply, the behavior is categorically different. Table 21 shows that all l-words are reliably selected over all boundary words. Moreover, there is a significant gap between the frequency of selecting l-words and the frequency of selecting boundary words. Little confusion exists and this was shown in a mean accuracy of 80% for the group. When we break down the selections by pairing (Table 22), we see the preference for l-words in all pairings. There remains a bias for trisyllabic l-words in that, for the trisyllabic-disyllabic pairings, the trisyllabic l-words are chosen more often than the disyllabic l-words. However, this “bias” does not extend to selecting any trisyllabic boundary words over disyllabic l-words.

There is no way to achieve these results using the lengthening parse or IV-feet so important to the interpretation of the other experiments. If lengthening were a primary means of segmentation, then the disyllabic words should be hard to discriminate. In a
lengthening parse, a disyllabic l-word is split into two feet. We saw this in 3A and 3B-Low. However, Table 22 shows the participants have no trouble separating disyllabic l-words from all boundary words. Similarly, they cannot only use IV-feet to score so well. IV-feet would include items such as kadu and tipa, and that should encourage them to select boundary words such as kaduda and tipabi. Only two words are selected less than kaduda in the trisyllabic-disyllabic pairings. The 3B-High group must be parsing the language in a very different way than the 3B-Low group (or 3A). There are two cues that can help achieve such a complete and accurate parse: the transitional probabilities between syllables and the rhythmic beat on each word onset. All of the other cues are too haphazard to trigger the resounding success of the 3B-High group. We first consider statistics.

Table 21. Experiment 3B-High Selected Words

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>tutipai</td>
<td>0.159</td>
</tr>
<tr>
<td>badi</td>
<td>0.139</td>
</tr>
<tr>
<td>gukadu</td>
<td>0.135</td>
</tr>
<tr>
<td>daku</td>
<td>0.131</td>
</tr>
<tr>
<td>pigaki</td>
<td>0.127</td>
</tr>
<tr>
<td>bibu</td>
<td>0.107</td>
</tr>
<tr>
<td>tipabi</td>
<td>0.060</td>
</tr>
<tr>
<td>bupi</td>
<td>0.040</td>
</tr>
<tr>
<td>gakitu</td>
<td>0.036</td>
</tr>
<tr>
<td>kaduda</td>
<td>0.036</td>
</tr>
<tr>
<td>kuba</td>
<td>0.020</td>
</tr>
<tr>
<td>digu</td>
<td>0.012</td>
</tr>
</tbody>
</table>
We previously suggested that SylProbs mark the end of a foot, perhaps because an increase in uncertainty is experienced as an increase in duration. A statistical foot, in the terms of this essay, would be perceived on each disyllabic l-word and the last two syllables of each trisyllabic l-word. Statistical feet then might explain the High group’s success on disyllabic l-words. It does not, however, explain the success with trisyllabic l-words. The statistical feet from the trisyllabic l-words *gukadu, tipabi,* and *pigaki* are *kadu, tipa,* and *gaki,* respectively. However, all of these statistical feet are in the trisyllabic boundary words as well, so participants should false alarm on those. They do not. This suggests that they have parsed the entire trisyllabic l-word – something we have not seen evidence for in the previous experiments.

To parse the entire trisyllabic l-word, some mechanism must exist to connect the first syllable of the trisyllabic with the final two syllables. Three different perceptions are considered in (54)-(56).
Example (54) shows two different feet, one heard using rhythm and one heard using SylProbs. Both would remain in memory and be used during the test. The results of the test argue against this representation. Again, the statistical feet should trigger false alarms, but they do not. Another possible metrical parse to consider is presented in (55).

In (55) the statistical foot is present on the last two syllables and then another foot is posited on the monosyllable gu where the rhythmic beat falls. This account has problems of stress clash with adjacent syllables and the familiar lack of predicted false alarms related to the statistical feet. Alternatively, the word gukadu could be parsed as having a stress on the 1st and 3rd syllables using two feet. This is a common English stress pattern such as in the words débonáir or céntralízed. This pattern applied to gukadu is presented in (56).

This parse has problems as well. There are few reasons to think the syllable du is stressed. It is part of a trisyllabic word and so will be shorter in duration than the following syllable 60% of the time. It has no beat occurring on it either. Finally, the beat does occur on the very next syllable, which would produce a stress clash. In short, (54) – (56) all have drawbacks as metrical parses of this stretch of speech.
If the participant can put all three syllables into a single foot, however, then the entire word will be heard. If we adopt an oscillator- or time-indexed metrical phonology (see Section 2.2.1), then we can allow for a three-syllable foot. In the oscillator model, we have one oscillator at the foot level entrained with a second oscillator at the syllable level. In binary feet, two syllables will be placed along the oscillator’s circle map, while in ternary feet, three syllables will be placed there. A representation of this is provided in Figure 14. We have the segments that will need to be spoken on the bottom level. These are aligned with the syllable-level oscillator. The syllable-level oscillator is in turn coupled with a foot-level oscillator. In the first section of the figure, the syllable oscillator goes through three periods for one period of the foot-level oscillator. This groups gu, ka, and du, into one foot, gukadu. In the second section, two syllables are placed within the foot, and then three syllables are again aligned for the third section.

<table>
<thead>
<tr>
<th>Foot</th>
<th>gukadu</th>
<th>daku</th>
<th>tutipa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syllable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>gu</td>
<td>ka</td>
<td>du</td>
</tr>
</tbody>
</table>

Figure 14. Two coupled oscillators coupled to produce three metrical feet.

This metrical parse then reveals all words in their totality as presented in the more traditional metrical grids of (57). The oscillator model of metrical phonology is not required here; its advantage is that it allows for the sorts of parses in (57) without multiplying the theoretical apparatus.
The metrical parse in (57) is supported by a beat occurring at each foot onset, as well as a perceived longer duration at the end of each foot due to the increase in uncertainty at l-word boundaries.

A little less than half of the participants in 3B appear to have found the complete parse of the language shown in (57), while the rest parsed the speech using the lengthening and IV-foot cues. Moreover, no one in 3A appears to have found the correct parse and remained with lengthening. Why? This is a difficult question to answer conclusively with the current data set. Earlier, we reviewed evidence that the lengthening parse was not quite as stable for the B language as it was for the A language. In 3A, many cues overlapped to cumulatively support the lengthening parse. For instance, the IV-feet *tipa* and *kadu* occurred as the 1st and 2nd syllables of a trisyllabic word. This reinforced the lengthening parse. Example (47) is copied below and slightly modified as (58) for convenience.

```
(58) ka du da di: gu: ti pa bi ku: ba: |
    | x     | x     | x     | x     | x     | x     | x   . |
Foot#  1  2  3  4  5  6
```

Feet #1 and #4 are both IV-feet and perceived due to the following perceived lengthening. In 3B, however, there was competing evidence indicating that the lengthening parse was not correct. In 3B, the IV-feet occur on the 2nd and 3rd syllables of the trisyllabic l-words. If participants were hearing these IV-feet, they would compete against the lengthening parse, since the lengthening parse tries to split the 2nd and 3rd syllables into different feet. If one keeps the 2nd and 3rd syllables together, one will also not divide following disyllabic l-words into different feet.
The participant is always searching for structure in the language. If a robust structure appears to be found with lengthening, such as in 3A, then the participant will remain with that parse. If the parse is not robust, then a participant might try an alternative structure – and find the correct answer (Gearhart et al., 2009; Gomez, 2002; Onnis, 2004).
CHAPTER 8. GENERAL DISCUSSION AND CONCLUSIONS

8.1. Overview of Findings

This dissertation considered two broad hypotheses: the Time Not Statistics hypothesis and the Time And Statistics hypothesis. Both hypotheses suggest that expectations for when in time a linguistic event will occur might be used to segment words within speech. In an artificial language such as the ones used in Saffran et al. (1996), the temporal pattern is an isochronous interval between word onsets, which could build an expectation for when the next onset will occur. The Time Not Statistics hypothesis made the bold claim that this temporal pattern was the way in which participants in Saffran et al. (1996) had succeeded in learning an artificial language, not the distribution of syllables. Such a claim was given credence by a study from Tyler and Johnson (2006) in which infants failed to segment a language with robust distributional cues but lacking the isochronous intervals between words. The Time And Statistics hypothesis simply proposed that the temporal expectation would be a cue as well as statistical distributions.

The design of the experiments was to assess these hypotheses by first testing adults on a Tyler and Johnson-style artificial language, and then modifying that language to restore isochronous word intervals. Restoring such word intervals required syllables of different durations. Unfortunately, the differing syllable durations were taken by many participants as final word lengthening, the phenomenon in which syllables at the ends of words (and prosodic groups) are longer in duration than other syllables in the word (or group). Through detailed analysis of the precise words that participants selected, there
was evidence that participants had indeed used this lengthening cue instead of the intended word onset interval.

Despite this, several patterns were discovered in the data that are of value. First, it is important to consider the time course of learning in an artificial language experiment. Some cues are available before others. Since speakers process speech incrementally and immediately, they will not wait for a cue such as statistical structure to emerge before attempting to find words. In the set of experiments here, cues such as final lengthening and aspirated voiceless stops were employed, even though they only provided a haphazard parse of the artificial language, since they are available immediately upon hearing. This suggests that future work should take the notion of temporal primacy into consideration when analyzing the process of segmentation.

Secondly, the results could not be explained without the concept of statistical learning. Of all the cues considered, none could entirely explain the participants’ word selections on forced choice tests, unless statistical cues were employed to parse the training language. This rejects the Time Not Statistics hypothesis. However, the units found through statistical learning were most consistent with metrical feet, not lexical entries. In other words, the statistical pattern highlighted prosodic not lexical boundaries. A possible motivation for this is the Information / Duration hypothesis, which suggests that increases in information processing, or uncertainty, can change how duration is subjectively perceived. Statistical calculations are experienced, not just calculated.

Third, we did find evidence that temporal patterning does affect speech perception, though not in the simple way intended in the experimental design. One group (in Experiment 3B-High) performed better than any other group, achieving a score of
80% correct, while other groups remained below 60%. This group’s behavior only matched a parse that included entire words, and the best cues in the artificial language for word onsets are the temporal pattern in conjunction with distributional cues. Moreover, the relatively low rate of learning in other conditions was most consistent with the presence of a competing segmentation cue. If participants group adjacent syllables due to their isochrony and do not find the pattern between words, this would explain much of the difficulty that participants had during the testing phase. All together, this supports the Time And Statistics hypothesis.

8.2. Revisiting the Nature of Statistical Learning

8.2.1. How do we calculate statistics?

In Chapter 1, we examined previous research that showed disagreement over what the precise statistical calculation might be for statistical learning. This is mostly assessed behaviorally by presenting a stimulus with one statistical pattern, not another, and seeing what is learned. However, there could be a measure of statistical uncertainty that is externally motivated.

As our first step, we can take a point made in Spivey (2007), who points out that any population of activated neurons can be understood as a probability distribution. At any point in time, a neural activation is partially in many different states. At syllable 1 in our experiments, the mind will make a prediction for the next syllable. We can call that prediction PD for Predicted Distribution. As syllable 2 arrives, it too will be represented in the mind. Ignoring the dynamic portions of this, it will be some other population code or some other Actual probability Distribution, AD. If the prediction of the next syllable is
correct, then PD will be very close to that of AD. If it is incorrect, then the distributions will be quite different.

Prediction creates a distance between the prediction and the actual datum, which can be represented as the distance between two probability distributions. The distance between two probability distributions \( p \) and \( q \) is defined as the Kullback-Leibler Divergence (\( D_{KL} \)) between \( p \) and \( q \), presented in (59).

\[
D(q||p) = \sum q(T) \log \frac{q(T)}{p(T)}
\]  

Levy (2008) explores a specific instance of the \( D_{KL} \) that is relevant for the current work. Levy proposes a surprisal measure of syntactic processing. In an incremental syntactic parser, a probability distribution for possible syntactic structures is built based upon the words heard so far. The distance between the parse after words \( S_{i,j} \) and the parse after the next word \( j \) is the distance between these two probability distributions or the \( D_{KL}(P_{i+1}||P_i) \). He then demonstrates that, given the extrasentential context remains constant from word \( i \) to word \( i+1 \), the distance between the two distributions is equivalent to the surprisal of word \( i+1 \). This is measured as the negative log probability of word \( i+1 \).

\[
\text{Surprisal} = D_{KL} = -\log P_i(w_{i+1})
\]  

An important thing to note about this measure of surprisal is that it is not simply one plausible calculation to be tested. Instead, if the assumptions hold, namely that any neural activation is a probability distribution and that the external context remains constant, then Levy’s surprisal measure is exactly the difficulty between moving from the previous state to the next state. Aslin et al. (1999) provide a list of possible probabilistic
measures that are compatible with their results. Levy provides an external motivation for believing surprisal to be the precise measure.

Let us walk through Experiment 1 using this measure of surprisal. After hearing syllable 1, the participant attempts to predict syllable 2. Notice that “prediction” in this case means moving into a neural space that approximates syllable 2. If syllable 2 as a datum matches this prediction, which it does if the participant has learned the language’s structure, then there is little surprisal between the prediction of syllable 2 and the actual syllable 2. It takes little effort to account for the new data. If their prediction is quite different from the real syllable 2, then they must quickly shift representations to the correct syllable. The greater the distance, i.e., the more surprising syllable 2 is, the more difficult such a transition will be.

From this logic, we can suggest a general definition for learning. The goal of learning is to minimize the distance between our predictions and the future data, i.e., the surprisal. The temporal dimension of surprisal should not be ignored, however. The difficulty is not simply in the distance we must go, but in how much time we have to get there. Intuitively, if we must move to a very different neural state space, but we have a long time to get there, the perceived difficulty may not be great. This is one motivation for incremental parsing, despite the garden paths that incremental parsing can lead us down. If we must move to an extremely different state space with rapidity, however, it will be quite difficult and require a large expenditure of energy. While it is beyond the scope of this work, one could map distance and time into an energy space such that the goal is to minimize the amount of energy needed to adjust for incorrect predictions.
8.2.2. Surprising information and duration in speech

Intriguingly, there is a documented relationship between information and duration in speech. Aylett and Turk (2004) performed a series of regressions for redundancy, duration, and prosodic structure in a large corpus of English. Redundancy in these terms is related directly to unpredictability and surprisal. If there is no new information in the next word, syllable, phoneme, etc., then that next unit is wholly redundant. Levy’s surprisal measures what is not predicted, i.e., not redundant. Aylett and Turk found an inverse relationship between redundancy and duration. If the speech signal was highly redundant (the surprisal was low), then durations were short. If the signal had little redundancy, durations were long. This also correlated with prosodic structure such that stressed syllables, for instance, both had longer durations and lower redundancy (more information).

Aylett and Turk therefore documented a relationship between surprisal and time on the production side. When surprisal was large, speakers produced longer durations. In Chapter 1, we reviewed behavioral studies of the subjective perception of time, which looks at a similar issue from the perception side (Pariyadath & Eagleman, 2007; Tse et al., 2004). What those studies documented was that when information processing was large, participants perceived it as an increase in duration. This is roughly the reverse of Aylett and Turk’s findings. We can map both the production and perception side into terms of movement through an energy space. When we expend lots of energy moving from one state to another in production, we increase the duration of our speech. Similarly, when we expend lots of energy moving from one state to another in perception, we experience it as an increase in duration. Eagleman and Pariyadath (2009) were able to
link the perception of time to energy devoted to neural computation. The essential idea is that neuronal response is suppressed during repetition. This can be viewed as an efficient form of encoding, since less metabolic energy is spent on redundant information. When new information arrives, the coding is not as efficient and increased energy must be spent.

If we take Aylett and Turk’s basic proposal that increased information leads to increased duration in speech production, it is natural to ask what the mechanism is. Here, we can call again upon our cherished oscillators. Chapter 2 briefly identified that a wealth of oscillator-based speech production models have been created in the last 10 years. Vousden et al. (2000) modeled the process of producing phonemes in the correct order. Later models, such as Tilsen (2009), Barbosa (2007), and Saltzman et al. (2008), move all the way from a phrase to segments or even articulatory gestures. All of these models take some sort of linguistic information and use oscillators to sequence and produce it in time. Increased durations of prosodic units are implemented by slowing the period of the oscillator at the correct point. Additionally, the oscillators are always constrained by the items they are presented with and their couplings to other oscillators. Therefore, if it takes more time to move from one neural state to another due to high surprise, this will slow the related oscillators either by not providing the required information quickly (for instance, the list of phonemes is not available for the syllable oscillator in time) or through a direct coupling such that the oscillator controlling articulatory gestures must slow its period to stay coupled.

Aylett and Turk (2004) provided corpus evidence that prosodically prominent syllables carried more information than others. Tilsen (submitted) indicates that more
neural effort is spent during stressed syllables as well. He employed a stop-signal paradigm in which the participant is instructed to immediately stop speaking at a signal. It took significantly longer for a participant to respond to the signal right before a stressed syllable than to a signal before an unstressed syllable. This suggests that more work is being done before stressed syllables, perhaps planning the articulatory gestures for the next foot, than before unstressed ones. In short, stressed syllables are both longer in duration, carry more information, and take more neural energy to plan.

8.2.3 Temporal packaging of sparse data

Through this work, we keep seeing a complex interaction between probability distributions and temporal patterns. Turning back to perception, oscillators might also be helpful in understanding another issue with statistical learning, namely what statistics to calculate.

The sort of artificial languages employed in this dissertation function like existence proofs. They do not directly mimic natural speech situations, but they are controlled enough to demonstrate a raw ability that might be put to use in real language situations. One manner in which the artificial languages used in experiments here differ from natural language is their extended and concentrated repetition. While the amount of speech that one hears in natural language contexts is enormous, it rarely happens in the sort of concentrated contexts that are employed here. All experiments in this dissertation repeated the same 6 words over and over for 7 minutes. While we may feel we have fallen into such situations in nightmarish conference presentations, this is quite rare in the real world.
If so and if we are still to use distributional information, we must have some way of capturing information in smaller doses and aggregating it over time. Perruchet and Vinter (1998) developed a PARSER model of segmentation in which various chunks of speech are parsed; chunks that are repeated are kept in memory; and chunks that are not repeated decay in memory. This model successfully segmented the language of Saffran et al. (1996), and it matches much of the analysis in this dissertation as well. Only in one group, 3B-High, did various partial parses get completely dropped for the complete and accurate model. In the other experiments, participants seemed to maintain various types of information into the test. Experiment 2A found evidence that participants perceived words based upon the aspiration of voiceless stops as well as statistical patterns. Even when there was an overlap, such as the l-word *kaduda* being split into *kadu* and *duda*, they did not merge the two representations, but kept them both. Grabbing partial chunks of information is a reasonable response to an environment in which information is sparsely distributed, but what “chunks” should one grab?

Various proposals have been offered. Newport and Aslin (2004) suggest gestalt principles, Yang (2004) suggests units determined by Universal Grammar, Endress et al. (2009) suggest dedicated mechanisms, such as edge detectors and repetition detectors, and Gomez (2006) suggests it will be determined by success at finding structure. Pacton and Perruchet (2008) argue instead that attention is both the necessary and sufficient condition. They implement a series of tasks that require participants to focus either on adjacent patterns or non-adjacent ones. After performing the task, participants would only display learning of whichever type of relation they had attended to during the task.
If we recall Large and Jones’s oscillator model again, we note that it is a model not simply of time perception but of attention. When an external stimulus couples only loosely with a dynamic oscillator, the attentional pulse is broadly distributed. When tightly coupled, however, attention has a sharp peak in tune with the oscillator’s phase. Similarly, following the expectancy scheme from Drake et al., (2000; see Chapter 2 for discussion), attention can be directed towards a particular oscillator that is strongly coupled with an external stimulus.

In Experiment 1, isochronous intervals existed between both syllables and word onsets. However, we saw no clear evidence that participants were able to find the inter-word onset. If they had found the full causal structure, the set of words, then they should have performed very highly during testing, such as we saw with the 3B-High group who achieved a score of 80%. Experiment 1 scored no higher than 58% (as a mean) despite having the same number of words and fewer syllables to recall than 3B-High. This is consistent with the notion that, while they were able to learn some structure through statistical distributions, they also were segmenting adjacent syllable pairs across word boundaries that were never dropped from memory. Imagine a simple oscillator with a period of 300 ms represented in Figure 15.

Figure 15 represents a syllable-level oscillator that perceives each syllable (ka, du, ti, etc.) as it arrives, so that each box for a syllable is a time stamp corresponding roughly to one period of the oscillator. Now, let us assume that the mind processes as a chunk any items which align together at a focal point of an oscillator, such as at phase(0). This is motivated by the notion from Large & Jones (1998) that periodic events in phase with an oscillator heighten attention. Each syllable will align with the following syllable in these
terms, since one oscillator is handling the syllables. The syllable *ka* aligns with phase 0 at period $P_i$, while the next syllable *du* aligns with phase 0 at period $P_{i+1}$. Adjacent items handled by the same oscillator will necessarily have attention directed towards them. This could construe the bias towards adjacent events in speech.

![Diagram](image)

**Figure 15.** A 300 ms. oscillator with syllabic input

Such chunking will sporadically group the correct syllables of the l-words of the artificial language, but it will also capture adjacent syllables that cross word boundaries. As intra-word chunks are repeated through the training period, they will be reinforced, while those that cross word boundaries will be reinforced less. This is similar to the PARSER model of Perruchet & Vintner (1998) and could account for the moderate 58% learning we saw in Experiment 1. The addition to the model from this dissertation is that the chunks could be based in part on rhythmic attention.

How might speakers move beyond reinforcement of adjacent syllables? What is needed is something that might provide hierarchical structure to the syllables. A word is more than two adjacent syllables. They have an order, *ka* is first and *du* second, and using English’ metrical phonology, *ka* is likely to be stressed while *du* is not, creating the English trochaic foot. A trochaic binary foot is defined as two syllables in which the stress aligns with the left side of the foot (Kager, 1999). Recall the stop-signal research of Tilsen (submitted), which indicated that a speaker could be planning the gestures for an
entire foot before a stressed syllable. One way to represent all this is to have the stressed syllable align with phase(0) of a foot oscillator, while the unstressed syllables align with less prominent positions (Figure 16).

Figure 16. A 600 ms oscillator coupled with a 300 ms oscillator

Figure 16 represents a single foot-level oscillator coupled with a single syllable-level oscillator. The syllable oscillator is represented twice because it oscillates twice over the period of the foot, once with the syllable *ka* and once with the syllable *du*. Phase (0) of the first cycle of the syllable oscillator is in phase with phase (0) of the foot-level oscillator, while phase (0) of the second cycle of the syllable oscillator is coordinated with a less prominent position. In this diagram, it aligns with phase (1/2).

A coupling of this sort provides a great deal more information about the linguistic structure than two sequential periods of the syllable oscillator alone. The first syllable *ka* is now designated as the first syllable of a larger structure. It also provides a concept of word edges. The syllable *ka* is the beginning edge of a single unit. In the same way that two syllables are processed as a chunk because they are handled by the same oscillator, two feet, as units, can now become a chunk. This allows calculations upon larger units as well as possibly informative edges (cf. Endress et al., 2009).

Chapter 2 suggested participants might be able to move from Figure 15 to Figure 16 by tracking a small number of possible models and finding the one with the most
structure. It appears that participants could not do this. Some greater cue of how the syllable and foot oscillators should align was needed. It only became possible in Experiment 3 when the isochrony of the syllables was reduced due to the difference in syllable duration and attention shifted to the higher-level foot oscillation.

The key proposal upon which we will end, then, is that periodic oscillators could serve as one mechanism for directing attention and thus for determining what statistics are calculated. Rhythmic attention acts as a filter to reduce the dimensionality of the speech stimulus. By providing a prosodic structure to speech, we can create structural units and edges through which to discover patterns.


