Table of Contents

Chapter 3: Online acquisition of Cairene Arabic word stress patterns over time.

Chapter 4: Is Less Really More? Answers from L2 Stress Acquisition

Chapter 5: Conclusion

Summary

Adult learners often experience difficulty in attaining fluency in all aspects of a second language (L2). The L2 acquisition of main word stress is especially challenging, and incorrect production may lead to significant miscommunication. Yet, relatively little research on L2 prosodic acquisition has been conducted. Additionally, most of this research focuses on English and other European languages, limiting our understanding of the acquisition of other stress systems. These studies contribute to this understanding by studying the L2 acquisition of the Cairene Arabic stress system.

Artificial Language Learning research methodology was combined with real Cairene Arabic input. Adult American English speakers with no previous experience of any other Semitic language were taught the Cairene Arabic stress system. In Chapter 3, patterns of acquisition similar to prior studies were observed, such as L1 transfer and overgeneralization, confirming that this is a viable method for studying beginning L2 acquisition. The experiment in Chapter 4 aimed to test the 'less is more' hypothesis in the domain of prosodic acquisition. The results in this experiment demonstrated that participants performed better when presented with random stimuli, as compared to initially limited stimuli which gradually increased in complexity. This showed that the 'less is more' hypothesis does not hold for the narrow domain of main stress acquisition.

These experiments were made possible by the use of Amazon Mechanical Turk (AMT), allowing a large number of participants to be tested quickly and cheaply. The studies in this thesis were conducted over the course of 4 days. However, online longitudinal experiments are difficult to conduct. While designing the above studies, a software solution called Longi was developed (Chapter 2), which simplifies the process significantly. It is hoped that the development of this software will enable more online longitudinal studies to be conducted.

Personal Declaration

I, Tamara Schembri, certify that the work in this dissertation entitled "Online Acquisition of Cairene Arabic Word Stress Patterns Over Time" is my original work and has not been previously submitted for a higher degree in any institution other than Macquarie University. I also certify that the thesis is an original piece of research and it has been written by me. Any help and assistance I have received during the course of my studies has been appropriately acknowledged. I also certify that all the information sources and literature used are indicated in this thesis. All the studies reported in this thesis gained approval from the Macquarie University Human Research Ethics Committee, reference number: 5201300185.

The material in this thesis is being prepared for publication in the publications listed below.

(1) Schembri, T., Johnson, M., & Demuth, K. In preparation. Longi: A Simple Automated System for Conducting Longitudinal Studies on Amazon Mechanical Turk. *Behavior Research Methods.*

(2) Schembri, T., Johnson, M., & Demuth, K. In preparation. Online acquisition of Cairene Arabic word stress patterns over time. *Applied Psycholinguistics*.

(3) Schembri, T., Johnson, M., & Demuth, K. In preparation. Is Less Really More? Answers from L2 Stress Acquisition. *Cognition*.

Acknowledgements

I am grateful to many people for their help and support while I was writing this thesis. Some gave critical feedback, or helped solving various problems. Others provided moral support in various ways. All of these people have provided essential contributions to this thesis. It will be impossible to thank all of them adequately, but I will do my best.

First of all, I must thank my supervisors, Katherine Demuth and Mark Johnson. The research project contained in this thesis first materialized following many discussions with my supervisors, suggesting new ideas and exploring many different possibilities. Many of the directions discussed did not materialize in the current thesis, but are explored in terms of future possibilities in the final chapter of this thesis. All of the ideas considered at one point or another contributed to make the overall grounding and direction of the current research far richer. Katherine Demuth suggested a number of connections I would not otherwise have made, such as the similarities between the acquisition of stress and tone systems, and the greater information value of longer rather than shorter words when acquiring these systems.

In the initial stages of preparing to conduct these experiments, there were many technical challenges to overcome. The stimuli in the current experiment were created using a speech synthesizer, and then manipulating the output in Praat. The expertise of Susan Lin was invaluable in helping me to segment syllables, work out how manipulations could be carried out and partially automated, as well as in providing encouragement when the techniques I was using just wouldn’t work. Once the stimuli were created, they didn’t always sound right. Audible pops, crackles and other artefacts from the manipulation techniques threatened to ruin the stimuli. Ivan Yuen helped to refine my techniques of extracting and manipulating my sound samples, so these issues could be resolved. However, this technique wasn’t the first one we tried! At the very beginning, we attempted to use syllables recorded by Kelly Miles, who spent

many hours recording syllables for me. However, this method caused issues when syllables with coda consonants when adjoined to following syllables with onset consonants. Many different techniques were attempted to resolve this issue, but it was John Pate who suggested that the speech synthesis program Festival might be the answer. Once all of the stimuli were created, they went through many rounds of approval with members of the Child Language Lab, who pointed out issues with specific stimuli and suggested many helpful solutions. For this, I thank all members of the Child Language Lab, but especially Susan Lin, Ivan Yuen, Nan Xu and Michael Proctor.

Technical challenges were also involved in creating the web-based interface used to conduct experiments in this thesis. Many thanks go to John Pate for fielding my many questions on Javascript, SQL, Linux and a whole host of other issues. And of course, thanks are due to the IT Department, who set up the virtual environment which made it possible to run these online experiments. Special thanks to Nan Xu for fielding my questions about experiment design and helping to ensure that the experiment was as well-balanced as possible while still using real-language stimuli. John Pate was instrumental in the design of the experiments contained in this thesis, suggesting a number of improvements and additions, and in the process giving me a much better grounding in statistics. Over the course of this thesis, I have had a number of in-depth conversations about statistics and mixed effects models with John Pate, Mark Johnson and Michael Proctor, which have deepened my understanding of the field, and allowed me to run the analyses contained in this thesis, which I would previously have been unable to do. Special thanks to Titia Benders for framing discussions about mixed effects models in a more intuitive way that made the fundamentals significantly easier to understand.

All of the members of the Child Language Lab have supported my research in some way, providing much needed feedback and moral support. My experience at Macquarie University would not have been the same without Ben Davies, Hui Chen, Sithembinkosi Dube, Qandeel

Hussain, Kiri Mealings, Carmen Kung, Elaine Schmidt, and Ping Tang. Similarly, my PhD journey would not have been the same without the year I spent as a visiting scholar at Amherst Umass under the guidance of John McCarthy and Joe Pater. I am grateful to Toni Borowsky for insisting that I go, and encouraging me along the way. Finally, special thanks must of course go to my husband, Peter Budziszewski, who propped me up throughout my PhD, and without whom this thesis would not have been possible.

Chapter 1: Introduction

This thesis investigates the acquisition of the Cairene Arabic stress system by adult American participants with no prior knowledge of Cairene Arabic or other Semitic languages. Adult learners often experience difficulty in attaining fluency in all aspects of a second language (L2); however, the acquisition of main word stress poses a particular challenge, with native-like attainment almost unattested in adults. Infants are sensitive to the rhythmic patterns of their language as early as 4 days old, and quickly learn to tune out non-native patterns. Infants' early acquisition of stress, during the first year of life, may help explain why adult prosodic acquisition is especially challenging. Nevertheless, adults' difficulties in correctly acquiring L2 stress patterns is problematic, as prosody has been shown to have a greater effect on intelligibility than segmental factors. Despite this, there has been relatively little research on L2 prosodic acquisition compared to other aspects of L2 acquisition, such as syntactic or morphological acquisition. Further, most previous research on the L2 acquisition of main word stress has focused on the acquisition of English and other European languages. As a result, we have a limited understanding of how other stress systems are acquired. The current set of studies aims to contribute to this understanding by examining the L2 acquisition of the Cairene Arabic stress system, which to our knowledge has not been previously studied. In addition, this thesis introduces and implements a set of methodological tools intended to facilitate future research into L2 acquisition. These are particularly useful for research into understudied languages, where participant recruitment may be difficult. In this chapter I briefly discuss the following background topics of relevance to these studies: a definition of stress; fundamental ways in which stress systems can differ; the first language (L1) acquisition of stress; L2 acquisition of stress; the effect of stress production on intelligibility; models of stress acquisition; a brief description of the content chapters contained in this thesis.

Stress Systems

Stress is the linguistic realization of rhythm. In stress languages, one syllable in a word is more prominent than others: this is the primary, or main stressed syllable. The phonetic correlates of stress are intensity, duration and pitch; however, the relative importance of each of these cues is language-specific (Hayes, 1995). Stress languages can broadly be divided into two types: fixed stress languages, in which stress is predictable, and free stress languages, in which stress is unpredictable and must be lexically specified. This is a distinction with important psycholinguistic consequences for the speakers of these languages. Speakers of fixed stress languages, in which invariant stress occurs at a word edge, lose the ability to discriminate suprasegmental contrasts (Dupoux, Peperkamp, & Sebastián-Gallés, 2001; Peperkamp & Dupoux, 2002); this is discussed in greater detail in the sections below. Stress languages can be further subcategorized under a number of different parameters; however, an understanding of the prosodic hierarchy is necessary to fully describe these parameters.

According to metrical stress theory, stress organization is hierarchical: individual segments are grouped into syllables (σ) and associated with a mora (μ), syllables are grouped into feet (Ft), and all the feet in a word are grouped together to form a Prosodic Word (PrWd or ω) (De Lacy, 1997). Figure 1, taken from de Lacy (1997), illustrates the organization of elements in the prosodic hierarchy with a phonetic representation of the word ‘onomastics’.

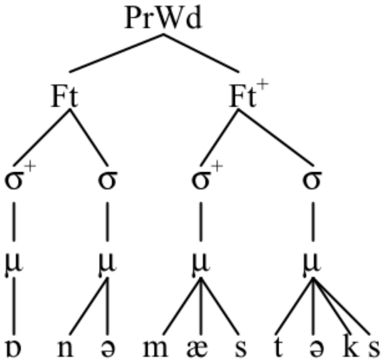


Figure 1: Prosodic organization of the word ‘onomastics’ (De Lacy, 1997).

Each element in the prosodic hierarchy contains a head, denoted by a plus sign (+). This is the strongest or most prominent element in each grouping. Each foot contains a single head syllable; in other words, a stressed syllable. The prosodic word may contain multiple feet; in the diagram above, it contains two feet, but only a single head foot. The head foot is the one which contains the main stressed syllable in the word. A foot which is not a head foot contains a head syllable with secondary, rather than primary, stress. Therefore, another definition of the main stressed syllable is that it is the head syllable of the head foot.

Stressed and unstressed syllables usually occur in an alternating pattern, demonstrated in Figure 1 above. Languages tend to avoid both adjacent stressed syllables, called clashes, as well as adjacent unstressed syllables, called lapses. There are two logically possible sequences of syllables that can create this alternating pattern: stressed-unstressed (also called strong-weak, left-headed or trochaic) or unstressed-stressed (also called weak-strong, right-headed or iambic). Figure 1 depicts a trochaic pattern. Stress languages tend to exhibit a strong preference for one of these two patterns, and can thus be categorized as trochaic or iambic stress systems. Trochaic stress systems are those in which feet are left-headed, while iambic stress systems are those in which feet are right-headed. Similarly, stress languages tend to assign main stress as closely as possible to either the left or right edge of the word. They can therefore be further categorized according to whether they are left- or right-headed on the level of the prosodic word. Figure 1 depicts a system which is right-headed on the level of the prosodic word, because main stress (on the syllable ‘mæs’) occurs in the rightmost foot. Note that this is despite the fact that feet in this language are trochaic, or left-headed: the headedness of different tiers is independent.

Another broad categorization of stress languages lies in the distinction between quantity sensitive and quantity insensitive stress systems. In quantity sensitive systems, syllables are categorized as light, heavy or superheavy. Heavy syllables attract stress over light syllables, and superheavy syllables attract stress over heavy syllables; in addition, some quantity sensitive languages may impose a minimum weight constraint on stressed syllables, disallowing stressed light syllables. Under a moraic analysis, light syllables are associated with a single mora, heavy syllables are associated with two moras, and superheavy syllables are associated with three moras. Languages may differ in the way segments are associated with moras, thus leading to differing classifications of syllables into weight classes. Syllables have an internal structure: they contain a vowel, also known as the nucleus, and may also contain an onset (the segment(s) preceding the nucleus) and a coda (the segment(s) following the nucleus). For example, the word ‘its’ has no onset, a nucleus ‘i’, and a coda ‘ts’. Onsets are not generally believed to contribute to weight, although counterexamples have been posited in the literature (Topintzi, 2006). Because of this, it is useful to make reference to the rime, which is the combination of nucleus and coda. For example, the word ‘cats’ has onset ‘c’ and rime ‘ats’. Weight systems come in two widely recognized types. One type of weight system allows vowels, but not consonants, to head a mora. As a result, syllables containing a long vowel are heavy, while all other syllables are not. Another system allows both vowels and consonants to head a mora. As a result, both syllables containing a long vowel, as well as syllables containing one or more coda consonants, are treated as heavy. In summary, all weight systems treat CV syllables as light, and CVV syllables as heavy. However, CVC syllables may be analyzed as either light or heavy, depending on the weight system. This is because languages differ in the moraic structure assigned to CVC syllables. The set of possible structures is illustrated in Figure 2 below.

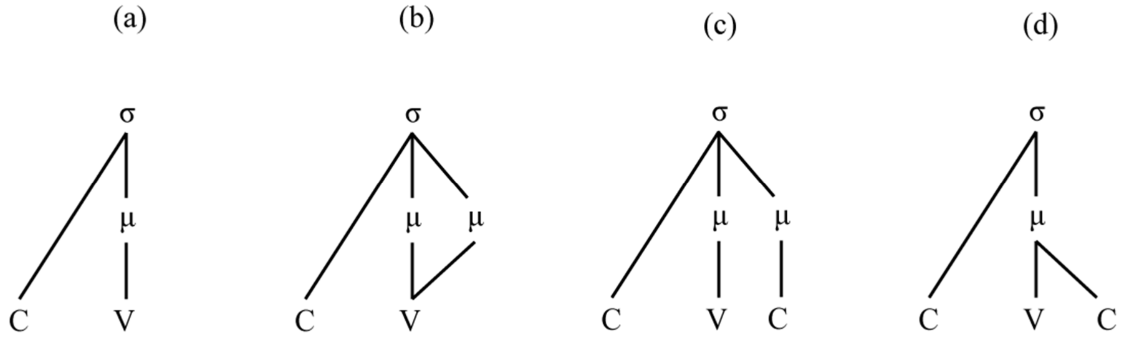


Figure 2: Light syllable (a); long vowel (b); heavy CVC syllable (c); light CVC syllable (d).

In this section, a number of different ways of classifying stress systems have been described: fixed and free stress languages; trochaic and iambic languages; systems which are right-headed and left-headed on the level of the prosodic head; quantity sensitive and quantity insensitive systems; systems in which only a long vowel can contribute to weight, and those additionally sensitive to coda consonants. These are classifications which have an early and significant effect on the first language (L1) acquisition of stress, which is described in the section below. Additional ways of classifying stress languages are described in the section on models of acquisition below.

L1 Acquisition of Stress

Infants are sensitive to the rhythmic properties of their native language from an early age. Newborn infants as young as 4 days old can use prosodic cues to distinguish speech in their native language from utterances in a foreign language (Mehler et al., 1988). This early ability to discriminate between languages is based on infants' ability to sort languages into a small number of rhythmic classes: those in which rhythm is based on the foot, such as English or Dutch; on the syllable, such as Italian or French; or on the mora, such as Japanese or Tamil (Dauer, 1987; Ramus, Dupoux, & Mehler, 2003). Newborns are able to discriminate between languages that belong to a different rhythmic class, but not between those that belong to the same rhythmic class (Nazzi, Bertoncini, & Mehler, 1998; Nazzi & Ramus, 2003; Ramus et al., 2003). Nazzi et al. (1998) presented newborns from French-speaking families with sentences from different languages which were low-pass filtered to remove segmental information. The infants were able to discriminate between English, a stress-timed language, and Japanese, a mora-timed language, but not between English and Dutch, which are both stress-timed languages. Ramus et al. (2003) demonstrated that these results hold even when acoustic manipulations are used to hold the intonation constant between utterances: in other words, infants are able to discriminate between languages using rhythmic properties alone. While newborns are able to discriminate between broad language classes, older infants begin to learn the language-specific prosodic properties of their native language. Nazzi, Jusczyk, & Johnson (2000) demonstrated that 5-month-old infants are able to discriminate between languages that belong to the same rhythmic class, but only when their native language, or one of its dialects, was among those presented. American infants were not able to discriminate between Dutch and German, even though both languages belong to their native rhythmic class. However, they were able to discriminate between a variety of English and another stress-timed language, as well as between American and British English. This result illustrates infants' growing sensitivity to the rhythmic patterns of their native language.

This sensitivity to the language-specific rhythmic properties of the native language results in diverging paths of development for infants from different language backgrounds (Bijeljac-Babic, Höhle, & Nazzi, 2016; Friederici, Friedrich, & Christophe, 2007; Höhle, Bijeljac-Babic, Herold, Weissenborn, & Nazzi, 2009; Skoruppa et al., 2009, 2013). Language-specific discrimination of stress patterns is evident in event-related brain potentials (ERP) in infants as young as 4 months old (Friederici et al., 2007). German bisyllabic words are usually stressed on the first syllable, while French words are stressed finally. French and German infants were exposed to initially and finally stressed bisyllabic words. The data from ERPs demonstrated a clear processing advantage for initially stressed words in German infants; and for finally stressed words in French infants. Similarly, behavioral experiments show the emergence of a trochaic preference between the ages of 4 and 6 months for German infants, while French 6-month-olds do not show a preference for either a trochaic or an iambic pattern (Höhle et al., 2009). Spanish 9-month-olds are able to discriminate between initially and finally stressed sets of words even when their segmental content is highly variable. However, French 9-month-olds can discriminate between initially and finally stressed sets of words only when they contain the same segmental content, such as *'pi.ma* and *pi.'ma* (Skoruppa et al., 2009). This indicates that French infants' failure to discriminate between different rhythmic patterns in segmentally dissimilar words is not due to an inability to perceive stress cues on an acoustic level. Instead, they are unable to process stress at a phonological level. This result from Skoruppa et al. (2009) demonstrates that stress deafness in speakers of fixed stress languages emerges at an early age. A further study by Skoruppa et al. (2013) more precisely determines that stress deafness emerges between the ages of 6 and 9 months old.

Infants learning a language with lexically contrastive stress such as English learn to pay close attention to stress cues, and to preferentially choose rhythmic cues when they are in opposition to other phonological or statistical cues. American infants display a clear preference for the predominant trochaic stress pattern between 6 and 9 months old (Jusczyk, Cutler, & Redanz, 1993). Infants are able to use this preference as a word segmentation strategy: at 7.5 months old, American infants are able to segment trochaic, but not iambic, words from fluent speech (Jusczyk, Houston, & Newsome, 1999). American 9-month-olds are sensitive to syllable weight, preferring stressed syllables which are heavy to those which are light (Turk, Jusczyk, & Gerken, 1995). A trochaic pattern with a light stressed syllable was contrasted with an iambic pattern with a heavy stressed syllable, testing whether infants had a stronger preference for trochaic patterns or heavy stressed syllables. They found that infants chose the word with a trochaic pattern, demonstrating a preference for the rhythmic properties of their native language over syllable weight. Similarly, when stress cues were designed to conflict with transitional probabilities, a statistical cue to word segmentation, 8-month-old infants were shown to preferentially rely on prosodic, rather than statistical information (Johnson & Jusczyk, 2001).

L2 Acquisition of Stress

Because infants quickly become attuned to the rhythmic patterns of their native language, learning to tune out non-native patterns, the second language acquisition of word stress in adulthood poses a significant challenge. A great deal of research on the acquisition of main word stress has focused on describing the stress assignment strategies of relatively homogenous groups: native speakers of the same L1, all learning a single L2. Jordanian Arabic learners of English reading out lists of real English words were shown to consistently produce words in accordance with their native stress rules (Anani, 1989). Egyptian Arabic learners of English, carrying out a similar task, consistently also used an L1 stress assignment strategy, correctly assigning stress only where the L1 and L2 stress patterns fell onto the same position in the word (Youssef & Mazurkewich, 1998). In these cases, L1 transfer was purely grammatical: learners directly used the fixed stress assignment rules of their L1. Where the L1 and L2 are more closely related, L1 transfer can also be lexical. English learners of German were shown to directly transfer stress patterns from the L1 when stressing closely related words, or cognates (Maczuga, 2014). However, the effect of cognates on stress assignment can be more subtle. For example, Baptista (1989) demonstrated that Brazilian learners of English most often assigned primary stress in English on the syllable which bears secondary stress in Portuguese cognates. Learners' non-native stress assignment strategies cannot always be entirely attributable to L1 transfer: learners sometimes produce interlanguage forms which are unlike either the native or target language. For example, Archibald (1992) found that Polish learners of English follow predictable patterns of error in stress placement. Some errors can be explained as L1 transfer: Polish has fixed penultimate stress, and words which should have final stress were incorrectly produced with penultimate stress. However, in other cases, learners produced antepenultimate or final stress on words which should have penultimate stress. As both English and Polish predict penultimate stress in these contexts, these errors cannot be attributed to transfer. Similarly, word stress is assigned right-to-left in both English and French; however, French learners of English assigned word stress from the left in a nonce word task (Pater, 1997). This is evidence of an interlanguage with characteristics found in neither the target nor the native language.

Another major research area concerns the perception of stress in nonce words in participants with a wide range of language backgrounds. Native speakers of languages with predictable stress perform poorly on stress perception (Altmann, 2006; Dupoux & Peperkamp, 2002; Dupoux et al., 2001; Peperkamp & Dupoux, 2002; Peperkamp, Vendelin, & Dupoux, 2010). However, native speakers of languages with non-predictable stress, such as Spanish, or without lexical stress, such as Thai (Altmann, 2006), perform at ceiling. Participants were required to learn two sets of minimal pairs, differing in either stress placement or place of articulation. For example, one minimal pair was [túku], associated with the key [1], and [túpu], associated with the key [2]. Long random sequences of each minimal pair were played to participants, who were asked to transcribe these as sequences of [1] and [2]. For example, the sequence [túku] [túku] [túku] [túpu] would be transcribed as [1, 1, 1, 2]. Dupoux et al. (2001) found that Spanish participants performed equally well on minimal pairs for stress placement or place of articulation; however, French participants performed significantly worse on stress compared to place of articulation. Peperkamp & Dupoux (2002) demonstrated that Finnish and Hungarian participants performed similarly to French participants. French has phrase-final accent, which can be described as word-final for the purpose of this experiment, while stress in Finnish and Hungarian is word-initial; therefore, stress deafness is independent of the position of word stress. However, although Polish has fixed penultimate stress, Polish participants did not exhibit stress deafness, performing more similarly to the Spanish participants. Polish stress is fixed, but it does not occur at a word edge. The authors hypothesize that this makes it more difficult for infants to extract this information in the first years of life; as a result, an abstract representation of stress is retained. Research on stress deafness implies that the L2 acquisition of main word stress is significantly more difficult for speakers of languages with predictable stress, who cannot correctly perceive the realization of stress in the input. However, the effects of stress deafness have been shown to be responsive to intensive training in an experimental setting (Carpenter, 2005). Although speakers of languages with predictable stress perform poorly in perception, these results are reversed in production. Altmann (2006) tested the same set of learners in both perception and production experiments using the same stimuli, finding that speakers of non-predictable languages perform best in production, while speakers of languages with lexical stress were less able to produce stress successfully.

Another body of research focusses on quantifying the effect of each independent factor affecting the successful perception and production of word stress in a specific language. Researchers can quantify these factors by examining the perception and production of different groups of advanced L2 learners. Groups of early and late Spanish (Guion, Harada, & Clark, 2004) and Korean (Guion, 2005) advanced learners of English were tested on the perception and production of word stress in English bisyllabic non-words. The results demonstrated that both perception and production were affected by a number of factors: phonological similarity to known words in either the native or target language, lexical class, and the presence or absence of long vowels and coda consonants. Age of acquisition was also a significant factor, with early learners patterning more closely with native speaker controls. For example, Spanish late learners ignored the presence of long vowels and coda consonants, while early learners, along with native controls, did not. As a result, the researchers concluded that the factors affecting perception and production do not have a consistent effect on all groups of learners, but influence early and late learners differently. Tremblay and Owens (2010) similarly examined the factors affecting successful L2 acquisition, comparing target-like advanced learners with less successful advanced learners. They asked French learners of English to produce non-words, with the aim of identifying consistent acoustic cues in successful learners' production. They found that the realization of stress in learners who assign stress incorrectly is associated with higher pitch, as compared to target-like learners and native controls.

Stress and Intelligibility

Stress information has a substantial effect on word recognition, lexical access and intelligibility. Word recognition can be facilitated by prior knowledge of stress patterns. Engdahl (1978) found that listeners were able to complete a sentence significantly faster when the stress pattern of the missing word was presented as a pattern of tones. Hirst & Pynte (1978) measured lexical access for words presented in uniform blocks, where words all contained the same stress pattern and number of syllables, finding that listeners responded significantly faster when compared to mixed blocks containing words of all stress patterns and syllable lengths. Even when vowel quality is controlled, such that unstressed syllables do not contain additional cues in the form of reduced vowels, stress information still affects word recognition (Cutler & Clifton, 1983). Conversely, stress patterns can result in false recognition. Participants in a nonce word experiment were presented with two sets of stimuli, and were asked to categorize the second set of words according to whether or not they had been previously presented. When stimuli were presented with different segmental information, but the same stress patterns as previously presented stimuli, they were incorrectly categorized as familiar (Robinson, 1977). Given that word recognition is significantly affected by the placement of stress, it is perhaps not surprising that incorrect use of stress by L2 learners has a significant effect on intelligibility. Non-standard syllable stress patterns can mislead native listeners into wrongly identifying words produced by a non-native speaker (Zielinski, 2008). In fact, prosodic errors can have a greater negative effect on intelligibility than segmental errors (Anderson-Hsieh, Johnson, & Koehler, 1992; Munro & Derwing, 1995). However, certain errors in stress placement are more likely to result in misperceptions than others. English native speakers are more likely to understand lexical stress errors that involve a leftward shift as compared to a rightward shift (Field, 2005; Lepage, 2015). This may be due to the predominance of initial stress in English (Cutler & Carter, 1987).

Models of Stress Acquisition

Dresher & Kaye (1990) proposed the first computational model of stress acquisition, based on metrical theory. They proposed eleven parameters to describe the stress system of a given language. Under this model, if a learner acquires the parameter settings for a given language, they will acquire the stress system. Dresher & Kaye's eleven parameters are often condensed into a system of eight parameters, excluding certain parameters which rarely come into play (Archibald, 1992; Van Der Pas & Zonneveld, 2004). This discussion follows Archibald (1992) and Van der Pas & Zonneveld (2004) in discussing only these eight parameters, as the remaining parameters are not relevant to the description of either the English or Cairene Arabic stress system. Four of these parameters have already been discussed in the section on stress systems above: feet can be left-headed (trochaic) or right-headed (iambic); stress systems can be left-headed or right-headed on the level of the prosodic word, resulting in main stress appearing closer to the left or right edge of the word; stress systems can be quantity sensitive or insensitive; in quantity sensitive systems, either a long vowel, or additionally a coda consonant, may contribute to weight. A number of additional parameters were proposed: 1) Feet can be constructed iteratively either left-to-right or right-to-left. 2) Feet can be either binary or unbounded. Unbounded feet can contain an unlimited number of syllables, and stress systems containing unbounded feet are always quantity sensitive. According to Prince (1990), feet may be binary under either a moraic or syllable analysis: that is, they must contain either two moras or two syllables. A foot containing a single heavy syllable is therefore binary under a moraic analysis, as it contains two moras. 3) A syllable may be designated extrametrical; that is, unable to attract stress. Archibald (1992) and Van der Pas & Zonneveld (2004) expand this parameter to allow for the designation of extrametrical consonants, as well as syllables. 4) Extrametrical syllables may fall on either the left or right edge of a word. In these stress systems, stress cannot fall on either the initial or final syllable, respectively.

Dresher and Kaye's model of stress acquisition is based on the Principles and Parameters model, in which binary parameters are either turned on or off throughout the course of acquisition. However, this does not provide the best model through which to analyze the differences between the English and Cairene Arabic stress systems. Both systems are trochaic; both are right-headed on the level of the prosodic word, with main stress occurring close to the right edge of the word; both are quantity sensitive; in both systems, coda consonants as well as long vowels contribute to weight; both systems contain binary feet; both systems contain extrametrical units at the right edge, at least under some analyses. In fact, Dresher & Kaye's model uncovers a single difference between the two stress systems: feet are constructed left-to-right in Cairene Arabic, and right-to-left in English. However, this list of similarities obscures some important differences between the two stress systems; these are briefly outlined in the current chapter, but described in greater detail in the content chapters of this thesis. Cairene Arabic is a predictable stress language, in which main stress assignment is entirely predictable based on phonological structure. English is a free stress language, in which main stress assignment is affected by phonological structure in a probabilistic manner, and must often be lexically specified. English and Cairene Arabic are both classified as quantity sensitive languages, but Cairene Arabic is arguably 'more' quantity sensitive than English. In Cairene Arabic, main stress assignment can be predicted without exception if the sequence of light (L), heavy (H) and superheavy (S) syllables is known. For example, a word with 3 heavy syllables (HHH) is always assigned penultimate stress, while a word with two heavy syllables and a final superheavy (HHS) is always assigned final stress. In English, weight has some effect on main stress assignment, but this is not entirely non-predictable. Such subtleties cannot be captured in a binary system in which the quantity sensitivity parameter is either 'on' or 'off', with no shades of grey in between. Similarly, CVV and CVC syllables are both treated as heavy in the English and Cairene Arabic systems.

However, in Cairene Arabic, CVV and CVC syllables contribute equally to weight in non-final position. Moreover, the pre-Optimality Theory analysis of Cairene Arabic represents final CVC syllables as underlyingly CV due to final consonant extrametricality; under this analysis, CVC and CVV syllables are equally likely to attract stress in all positions. Conversely, in English, CVV syllables contribute to stress assignment to a far greater extent than do CVC syllables (Guion, et al., 2003; Guion et al., 2004; Guion, 2005). Once again, this difference between the two stress systems cannot be simply stated in terms of principles and parameters: it is not a binary difference, but one of degree.

Optimality Theory (OT) provides a framework which can better account for these non-binary distinctions. Constraints in OT may be active in a given language, yet still violable, such that winning candidates may obey a given constraint in some contexts but not others. Both English and Cairene Arabic are quantity sensitive. However, this is the main driver of stress assignment in Cairene Arabic, while in English, quantity sensitivity does not have as great an effect on stress assignment. In OT, these differences can be explained through constraint ranking and interaction. In Cairene Arabic, the constraints which govern weight-driven stress assignment are highly ranked. In English, these constraints are ranked high enough to have some effect on stress assignment, but are outranked by competing constraints in many contexts. The learnability of stress systems has been extensively studied within the framework of Optimality Theory (Apoussidou, 2007; Tesar, 1997; Tesar & Smolensky, 1998). This is because input/output faithfulness interactions are not a concern in the acquisition of stress systems. Input and output forms differ only in the assignment of stresses. Therefore, a learning model need not be concerned with a mechanism to deduce the underlying form from the speech signal: the underlying form is simply the output form without any stress. Tesar (1997) proposed a learning model to describe all possible stress languages, positing twelve universal constraints. These fulfil a similar function to Dresher & Kaye's (1990) parameters; however, rather than setting constraints on or off, learners must acquire the rankings between them. These proposed universal constraints, with some modifications, were used in the analysis of Cairene Arabic contained in the content chapters of this thesis, and are discussed below.

The constraint WSP states that all heavy syllables must be stressed; any unstressed heavy syllables incur a violation of this constraint. The constraint PARSE-SYLLABLE ensures that every syllable must be footed; any unfooted syllable incurs a violation of this constraint. The constraint LAPSE-FT was proposed by de Lacy (2002) to ban two adjacent unfooted syllables. Under a system containing both of these constraints, two adjacent unfooted syllables would incur a single violation of LAPSE-FT, but two violations of PARSE-SYLLABLE. These constraints therefore fulfil a similar function. The constraint LAPSE-FT was used in the analysis of Cairene Arabic; however, an analysis using only PARSE-SYLLABLE would not change any constraint rankings or interactions. The OT analysis of Cairene Arabic in the content chapters in this thesis discusses only active constraints. The effect of the constraint PARSE-SYLLABLE is to rule out large sets of candidates which are not considered in this analysis; for example, output candidates with no feet and no stresses. The matching constraints MAIN-RIGHT and MAIN-LEFT correspond to a single parameter in Dresher & Kaye's model. These determine whether stress tends to occur closer to the left or right edge of the word. These are gradient alignment constraints assigning a violation for each constituent between the relevant foot-edge and word-edge. McCarthy (McCarthy, 2003) argued against the use of gradient constraints in OT, arguing instead for categorical variants. Therefore, the constraint ENDRULE-R, the categorical variant of MAIN-RIGHT, is used in the current analysis. This assigns a single violation for any non-final foot which contains a main stressed syllable. The matching constraints WORD-FOOT-LEFT and WORD-FOOT-RIGHT require either the left or right edge of the word to coincide with a foot. WORD-FOOT-LEFT incurs a single violation for any initial unfooted syllable; WORD-FOOT-RIGHT incurs a single violation for any final unfooted syllable. In the current analysis, I use the equivalent ALIGN-L (Kager, 2001, 2005). The matching constraints ALL-FEET-RIGHT and ALL-FEET-LEFT are gradient alignment constraints which require feet to coincide with either the left or right edge of the word. These constraints have been shown to have multiple undesirable consequences; their effects can be replicated through the constraint family LAPSE, as well as the constraints ALIGN-L and ALIGN-R (Kager, 2001, 2005; McCarthy, 2003). Therefore, they are not used in the current analysis. The matching constraints IAMBIC and TROCHAIC are used to ensure that feet are either iambic or trochaic respectively. As Cairene Arabic is a trochaic system, the constraint TROCHAIC is used in the current analysis; it incurs a violation for each foot which is not trochaic. The matching constraints NONFINAL and NONINITIAL are used to create the effect of extrametricality on the right and left edge respectively. The constraint NONFIN is used in the current analysis. In Tesar's (1997) formulation, this states that final syllables should not be footed. In the current analysis, Prince & Smolensky's (2002) more stringent definition is used, banning main stress in the final foot. A few additional constraints used in the current analysis were not included in Tesar's framework. The constraint FTBIN requires feet to be binary. Tesar does not make explicit reference to this constraint; however, he only considers candidates which obey this constraint. WSP-CVV is an ad-hoc constraint. This is shorthand for a complex set of constraint rankings which ensure that long vowels are always stressed in the surface form, regardless of position; for a full analysis of these constraints, please see McCarthy (2005). The constraint WSPμμμ is a scalar version of the constraint WSP, and refers to superheavy syllables rather than heavy syllables. Tesar's input strings contained only light and heavy syllables, and so did not make use of this constraint.

Organization of Thesis

The first content chapter (Chapter 2) is titled ‘Longi: A Simple Automated System for Conducting Longitudinal Studies on Amazon Mechanical Turk’. This introduces Longi, a piece of custom software I designed which makes it easier to run longitudinal, or multi-session experiments on Amazon Mechanical Turk. Longitudinal studies enable researchers to investigate a broader range of research questions than would be possible with a cross-sectional design. Traditional longitudinal studies are costly and time-consuming; as a result, they are relatively uncommon. Conducting such research online is significantly cheaper and faster; however, online longitudinal studies are rarely carried out. This is because of the significant technical challenges involved in designing such studies. Longi automates many of the tasks involved, such as publishing new experiment sessions on a set schedule and sending out reminders to participants to reduce attrition rates. This software enabled the research described in this thesis, allowing experiments to be held on a larger scale than would have otherwise been possible.

The second content chapter (Chapter 3) is titled ‘Online acquisition of Cairene Arabic word stress patterns over time’. This experiment combined the methodology of Artificial Language Learning research with real-language input, extracted from the LDC Colloquial Egyptian Arabic lexicon. Participants were adult American English speakers with no previous experience with any other Semitic language. The experiment was run using custom Javascript software on Mechanical Turk, and was held over the course of 4 days. Participants were taught the Cairene Arabic stress system, which is highly complex. The use of Longi enabled acquisition of such a complex system, which would not be possible over a single session. The aim was to quantify the factors affecting participants’ acquisition. Evidence was found of acquisition patterns seen in the L2 acquisition of other stress systems, such as L1 transfer and overgeneralization. In addition, differential rates of acquisition for two sets of constraint rankings were demonstrated. Participants performed significantly better on words in the constraint group WSPμμμ, which demonstrated the ranking WSPμμμ >> NONFIN than on words in the constraint group WSP-CVV, which demonstrated the ranking WSP-CVV >> NONFIN. More broadly, this experiment demonstrated how research into beginning L2 acquisition can be facilitated through the use of online experimentation and crowdsourcing technology. This is especially true for research into understudied languages: as participants begin the experiment with no prior knowledge of the target language, recruitment for such studies becomes significantly easier. Our current understanding of L2 acquisition is disproportionately drawn from English and other European languages; as a result, our knowledge of the general mechanisms underlying acquisition are necessarily limited. These methods are intended to facilitate research in this area, increasing the availability of data for the acquisition of a wider range of languages. The current experiment is a first step in this direction, providing data on the L2 acquisition of Cairene Arabic, which to our knowledge had not been previously studied.

The third content chapter (Chapter 4) is titled 'Is Less Really More? Answers from L2 Stress Acquisition'. This experiment tests the 'Less is More' hypothesis (Newport, 1990), which states that L1 acquisition is facilitated by children's limited working memory capacity, and that acquisition in adults can be improved by limiting the input initially provided. Researchers have tested this hypothesis in multiple domains, with mixed results. Some studies have demonstrated a 'less is more' effect, while others have shown the opposite result. The current paper compares participants' acquisition of the Cairene Arabic stress system under two conditions. In the experiment described in Chapter 3, the stimuli presented to participants gradually increased in complexity. In the experiment in Chapter 4, stimuli were presented randomly, such that stimuli of all levels of complexity could be encountered by participants at any point in the experiment. Each participant was presented with a unique ordering, selected randomly by the software on initialization. The two experiments were identical in all respects other than presentation order. This allowed for a direct comparison of adults' acquisition of initially limited input, versus input which immediately contains the full range of complexity. This is the first direct comparison of this kind in the domain of prosodic acquisition. Participants in the random experiment (Chapter 4) outperformed those in the limited input experiment (Chapter 3) overall; additionally, performance was better on all metrics measuring participants' underlying knowledge of aspects of the prosodic structure of Cairene Arabic. Given these results, it appears that, for the narrow domain of word stress acquisition, the 'Less is More' hypothesis does not hold. Additionally, the experiment in Chapter 4 demonstrates that the methodology used throughout this thesis, combining online experimentation, crowdsourcing and real-language data, can be applied to more traditional ALL research questions, as well as pure research into L2 acquisition. The final chapter (Chapter 5) summarizes all results, discusses limitations of the current research, and explores future directions and implications of this research.

References

Altmann, H. (2006). *The perception and production of second language stress: A cross-linguistic experimental study*. University of Delaware.

Anani, M. (1989). Incorrect stress placement in the case of Arab learners of English. *IRAL-International Review of Applied Linguistics in Language Teaching*, *27*(1), 15–22.

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The Relationship Between Native Speaker Judgments of Nonnative Pronunciation and Deviance in Segments, Prosody, and Syllable Structure. *Language Learning*, *42*(4), 529–555.

Apoussidou, D. (2007). *The learnability of metrical phonology*. LOT.

Archibald, J. (1992). Transfer of L1 parameter settings: Some empirical evidence from Polish metrics. *Canadian Journal of Linguistics/Revue Canadienne de Linguistique*, *37*(3), 301–340.

Baptista, B.O. (1989). Strategies for the Prediction of English Word Stress. *International Review of Applied Linguistics* 27, 1, Feb,1-14.

Bijeljac-Babic, R., Höhle, B., & Nazzi, T. (2016). Early prosodic acquisition in bilingual infants: the case of the perceptual trochaic bias. *Frontiers in Psychology*, *7*.

Carpenter, A. C. (2005). Acquisition of a natural vs. unnatural stress system. In *Proceedings of the 29th annual Boston University Conference on Language Development* (pp. 134–43).

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech & Language*, *2*(3–4), 133–142.

Cutler, A. & Clifton, C. E. (1983). Lexical stress effects on phonetic categorization in auditory word perception. Paper presented to the Tenth International Congress of Phonetic Sciences, Utrecht.

Dauer, R. M. (1987). Phonetic and phonological components of language rhythm. In *Proceedings of the XIth International Congress of Phonetic Sciences* (Vol. 5, pp. 447–450). Tallinn.

De Lacy, P. (1997). *Prosodic categorisation*. University of Auckland.

De Lacy, P. V. (2002). *The formal expression of markedness*. University of Massachusetts Amherst.

Dresher, B. E., & Kaye, J. D. (1990). A computational learning model for metrical phonology. *Cognition*, *34*(2), 137–195.

Dupoux, E., & Peperkamp, S. (2002). Fossil markers of language development: phonological “deafnesses” in adult speech processing. *Phonetics, Phonology, and Cognition*, 168–190.

Dupoux, E., Peperkamp, S., & Sebastián-Gallés, N. (2001). A robust method to study stress "deafness." *The Journal of the Acoustical Society of America*, *110*(3), 1606–1618.

Engdahl, E. (1978). Word stress as an organizing principle for the lexicon. Papers from the Fourteenth Regional Meeting, Chicago Linguistic Society.

Field, J. (2005). Intelligibility and the Listener: The Role of Lexical Stress. *TESOL Quarterly*, *39*(3), 399–423.

Friederici, A. D., Friedrich, M., & Christophe, A. (2007). Brain responses in 4-month-old infants are already language specific. *Current Biology*, *17*(14), 1208–1211.

Guion, S. G. (2005). KNOWLEDGE OF ENGLISH WORD STRESS PATTERNS IN EARLY AND LATE KOREAN-ENGLISH BILINGUALS. *Studies in Second Language Acquisition*, *27*(4), 503–533.

Guion, S. G., Harada, T., & Clark, J. J. (2004). Early and late Spanish–English bilinguals' acquisition of English word stress patterns. *Bilingualism: Language and Cognition*, *7*(3), 207–226.

Hayes, B. (1995). *Metrical stress theory: Principles and case studies*. University of Chicago Press.

Hirst, D. J. & J. Pynte (1978). Gender, stress and tone: arbitrary features in the organisation of the lexicon. *Sigma* 3: 73-87.

Höhle, B., Bijeljac-Babic, R., Herold, B., Weissenborn, J., & Nazzi, T. (2009). Language specific prosodic preferences during the first half year of life: Evidence from German and French infants. *Infant Behavior and Development*, *32*(3), 262–274.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*(4), 548–567.

Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Infants' Preference for the Predominant Stress Patterns of English Words. *Child Development*, *64*(3), 675–687.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, *39*(3), 159–207.

Kager, R. W. J. (2001). Rhythmic directionality by positional licensing.

Kager, R. W. J. (2005). Rhythmic licensing theory: an extended typology. In *Proceedings of the third international conference on phonology* (pp. 5–31). Seoul National University.

Lepage, A. (2015). *The contribution of word stress and vowel reduction to the intelligibility of the speech of Canadian French second language learners of English*. Université Laval.

Maczuga, P. S. (2014). *Production of German L2 Stress by Native Speakers of English*. University of Calgary.

McCarthy, J. (2003). OT constraints are categorical. *Phonology*, *20*(1), 75–138.

McCarthy, J. J. (2005). The length of stem-final vowels in Colloquial Arabic. *Perspectives on Arabic Linguistics XVII–XVIII*, 1–26.

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

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73–97.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 756.

Nazzi, T., Jusczyk, P. W., & Johnson, E. K. (2000). Language Discrimination by English-Learning 5-Month-Olds: Effects of Rhythm and Familiarity. *Journal of Memory and Language*, *43*(1), 1–19.

Nazzi, T., & Ramus, F. (2003). Perception and acquisition of linguistic rhythm by infants. *Speech Communication*, *41*(1), 233–243.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*(1), 11–28.

Peperkamp, S., & Dupoux, E. (2002). A typological study of stress "deafness." *Laboratory Phonology*, *7*, 203–240.

Peperkamp, S., Vendelin, I., & Dupoux, E. (2010). Perception of predictable stress: A cross-linguistic investigation. *Journal of Phonetics*, *38*(3), 422–430.

Pater, J. (1997). Metrical parameter missetting in second language acquisition. *Language acquisition and language disorders*, *16*, 235-262.

Prince, A. (1990). Quantitative consequences of rhythmic organization. *CLS*, *26*(2), 355–398.

Prince, A. S. (1985). Improving tree theory. In *Annual Meeting of the Berkeley Linguistics Society* (Vol. 11, pp. 471–490).

Prince, A., & Smolensky, P. (1993/2002). Optimality Theory: Constraint Interaction in Generative Grammar (Technical).

Ramus, F., Dupoux, E., & Mehler, J. (2003). The psychological reality of rhythm classes: Perceptual studies.

Robinson, G. M. (1977). Rhythmic organization in speech processing. *Journal of Experimental Psychology: Human Perception & Performance*, 3: 83-91.

Skoruppa, K., Pons, F., Bosch, L., Christophe, A., Cabrol, D., & Peperkamp, S. (2013). The Development of Word Stress Processing in French and Spanish Infants. *Language Learning and Development*, *9*(1), 88–104.

Skoruppa, K., Pons, F., Christophe, A., Bosch, L., Dupoux, E., Sebastián-Gallés, N., … Peperkamp, S. (2009). Language-specific stress perception by 9-month-old French and Spanish infants. *Developmental Science*, *12*(6), 914–919.

Tesar, B. (1997). An iterative strategy for learning metrical stress in Optimality Theory. In *Proceedings of the 21st Annual Boston University Conference on Language Development* (pp. 615–626).

Tesar, B., & Smolensky, P. (1998). Learnability in optimality theory. *Linguistic Inquiry*, *29*(2), 229–268.

Topintzi, N. (2006). *Moraic Onsets*. University of London.

Tremblay, A., & Owens, N. (2010). The role of acoustic cues in the development of (non-) target-like second-language prosodic representations. *The Canadian Journal of Linguistics / La Revue Canadienne de Linguistique*, *55*(1), 85–114.

Turk, A. E., Jusczyk, P. W., & Gerken, L. (1995). Do English-Learning Infants use Syllable Weight to Determine Stress? *Language and Speech*, *38*(2), 143–158.

Van Der Pas, B., & Zonneveld, W. (2004). L2 parameter resetting for metrical systems (An assessment and a reinterpretation of some core literature). *Linguistic Review*, *21*(2), 125–170.

Youssef, A., & Mazurkewich, I. (1998). The Acquisition of English Metrical Parameters and Syllable Structure by Adult Native Speakers of Egyptian Arabic (Cairene Dialect). In *THE GENERATIVE STUDY OF SECOND LANGUAGE ACQUISITION, Flynn, Suzanne, Martohardjono, Gita, & O'Neil, Wayne [Eds], Mahwah, NJ: Lawrence Erlbaum Associates, 1998, pp 303-332*.

Zielinski, B. W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System*, *36*(1), 69–84.

Chapter 2: Longi: A Simple Automated System for Conducting Longitudinal Studies on Amazon Mechanical Turk

ABSTRACT

The last few years have seen a significant increase in the popularity of online platforms such as Amazon Mechanical Turk (AMT). However, researchers have largely confined themselves to simple designs. In particular, there have only been a small number of longitudinal studies carried out on the platform. This is in large part due to the significant technical difficulties involved in designing such a study. We argue that AMT is an excellent platform for longitudinal designs and learning studies held over multiple sessions; it enables designs which might otherwise not be practical. We aim to facilitate the development of such studies by introducing Longi, a script which automates many of the common tasks associated with publishing longitudinal studies on AMT. Longi is available online at http://github.com/tschembri/Longi.

Introduction

Researchers in the behavioral sciences have increasingly begun to rely on web-based experiments due to the numerous benefits of conducting research online. These include large scale data collection (Kramer, Guillory, & Hancock, 2014), access to a large and diverse subject population (Buhrmester, Kwang, & Gosling, 2011), and significant cost and time savings (Mason & Suri, 2012). A number of platforms for web-based experiments are available, such as SurveyMonkey, Crowdflower and CrowdGuru. However, the most prominent is Amazon Mechanical Turk (AMT), as it allows researchers to access a large, pre-existing population of potential participants who have been extensively studied, as well as tools such as qualifications, which can be used to control the available participant pools for a given experiment. Consequently, the current article will focus on AMT exclusively.

AMT is used to prepare materials for offline studies (Fort, Martin, & Peperkamp, 2015), and as a platform to conduct experiments (Culbertson & Adger, 2014; Tily, Frank, & Jaeger, 2011). While various types of experimental design can be replicated online, reports by Gureckis et al. (2015) and Simcox & Fiez (2014) conclude that the majority of web-based behavioral research is limited to survey-like designs. This is in large part due to the significant technical difficulties involved in developing online experiments. In response, a number of researchers have designed frameworks to facilitate the development of specific experimental designs. Among other examples, these applications allow researchers to design experiments with individual trials organized into blocks (de Leeuw, 2014), reaction time measurements (Simcox & Fiez, 2014), and real-time, synchronous interactions between participants (Hawkins, 2015). Similarly, the current paper presents Longi, an open-source script designed to simplify the task of developing longitudinal studies on Amazon Mechanical Turk. Longi can be downloaded from github (http://github.com/tschembri/Longi).

Longitudinal Studies

Longitudinal studies enable researchers to establish relationships between variables which cannot be uncovered through a cross-sectional design (Bauer, 2004). As a result, they allow researchers to investigate a broader range of research questions.

Longitudinal studies on AMT

Relatively few studies have used AMT to conduct research over multiple sessions. Researchers have collected survey responses after one-week (Shapiro, 2013; Carr, 2014), a year (Chandler, Mueller and Paolacci, 2014), and two, four, eight, and thirteen months (Daly & Nataraajan, 2015); estimated test-retest reliability over a three-week period (Buhrmester et al., 2011; Holden et al., 2013); conducted experience sampling daily over 14 consecutive days (Boynton & Richman, 2014), and twice daily over 10 consecutive days (Lanaj, Johnson, & Barnes, 2014). Daly & Nataraajan (2015) argue that the lack of longitudinal designs on AMT is due to the significant technical difficulties involved in conducting such studies online.

Learning longitudinally

Learning studies often take place over multiple sessions. Participants have been taught: an artificial grammar, over 7-10 days (Hudson, Kam & Newport, 2005); to distinguish between categories, over 5 days (McKinley & Nosofsky, 1995); to recognize visual stimuli, over 4 days (Standing, Conezio, & Haber, 1970); to navigate an environment, over 5 days (Foreman, Stanton, Wilson, & Duffy, 2003). When information is distributed over time, it can be better acquired. A meta-analysis of this effect (Cepeda, Pashler, Vul, & Wixted, 2006) concludes that increasing the time between learning sessions improves retention; increasing the time between the final learning session and a test has a similar effect. Sleep consolidation additionally improves performance when learning sessions are timed appropriately; this effect occurs only in studies of generalized, rather than rote, learning (Fenn, Margoliash, & Nusbaum, 2013). As such, learning studies are particularly well suited to a longitudinal design.

Learning studies on AMT

Learning studies are commonly found on AMT. In some studies, participants are taught specific material, and their ability to generalize from the data is tested. Most commonly, this involves participants being taught an artificial language. For example, participants have learned: to combine nouns and classifiers (Culbertson & Adger, 2014; Culbertson & Wilson, 2013; Tily et al., 2011), a phonological identity effect (Gallagher, 2013; Linzen & Gallagher, 2014), the collocational behavior of novel verbs (Paciorek & Williams, 2015), non-adjacent dependencies (Enochson, 2015), word meanings (Horowitz & Frank, 2012; Frank & Goodman, 2014), to identify referents through pragmatic inference (Vogel, Emilsson, Frank, Jurafsky, & Potts, 2014).

Other studies investigate the process of learning: researchers have examined the effect of variable retrieval practice on learning (Maas, Pavlik and Hua, 2015), the prevalence and effectiveness of participants' learning strategies in discovering causal relationships (Rottman, 2014), the effect of individual vs. collective knowledge on participants' choice of social tags (Cress, Held, and Kimmerle, 2012), the factors influencing participants' ability to learn a new social norm (Hareli et al., 2015).

Learning longitudinally on AMT

Despite the prevalence of learning studies on AMT, and the suitability of such studies for a longitudinal design, very few are held over multiple sessions. Participants have been taught to use novel authentication methods, over a period of 7 days (Forget, Chiasson, & Biddle, 2012), to complete a variety of language tasks, over 3 sessions (Enochson, 2015), a complex stress system using Longi, over 4 days (Schembri, Johnson, and Demuth, 2016a, 2016b); while Zhu, Dow, Kraut, and Kittur (2014) examined the effect of mentoring on learning, over 2 sessions.

AMT: Advantages for Longitudinal Studies

There are a number of benefits involved in conducting longitudinal studies on AMT, rather than offline. These may enable researchers to carry out experiments which would otherwise not have been feasible.

Time and cost savings

Longitudinal studies are significantly more expensive and time-consuming to conduct than cross-sectional designs. As a result, fewer subjects can be included for a given budget than in a single-stage study. AMT allows researchers to recruit large numbers of participants at a significantly reduced cost (Mason & Suri, 2012). This can enable the design of longitudinal studies for which funds would not otherwise be available. Large amounts of data can be collected extremely rapidly. This can save considerable time and money on participant recruitment. The speed of data collection is affected by the desired geographical location of participants: Demmet et al. (2015) collected data from 505 US participants in under 2 days, 505 Indian participants in 11 days, and 118 participants from other countries in over 30 days.

Geographical mobility

Traditional longitudinal studies are hampered by participants who do not remain in the same geographical area (St Pierre, 1980; van Weel, 2005). Even where study designs do not exclude participants who move during the experiment, these individuals may be harder to locate for subsequent rounds. AMT eliminates many of these concerns, as participants' geographical mobility no longer affects the researcher's ability to contact them.

Specific or hard-to-reach populations

Longitudinal studies can be adversely affected by failing to include hard-to-reach respondents (Odierna & Schmidt, 2009). AMT allows researchers to target specific participant profiles, and makes it significantly easier to access hard-to-reach populations (Smith, Sabat, Martinez, Weaver, & Xu, 2015). Studies have targeted individuals with physical disabilities (Tenenbaum, Byrne, & Dahling, 2014), individuals with psychiatric symptoms (Shapiro, Chandler, & Mueller, 2013), adult cancer survivors (Carr, 2014), LGB individuals (Vaughn, Cronan, & Beavers, 2015), fathers, who are severely underrepresented in clinical child and adolescent research (Schleider & Weisz, 2015; Parent et al., 2015), pregnant women (Arch, 2014), and veterans of Operation Iraqi Freedom (Lynn, 2014). Although there are some concerns about the accuracy of self-reported demographic data, researchers have developed several methods of ensuring truthful reporting (Smith et al., 2015; Chandler & Shapiro, 2016).

Experiment designs

Researchers who wish to conduct longitudinal experiments on AMT are able to easily replicate many lab-based designs, and can expect to achieve similar results (Casler, Bickel, & Hackett, 2013; Goodman, Cryder, & Cheema, 2013; Holden et al., 2013). Additionally, the online platform enables novel designs which may be impractical to achieve in offline longitudinal studies. For designs which feature learning over time, adaptive training techniques (Raybourn, 2007; Stacey et al., 2010) can be used to modify the difficulty of test items based on past responses. Researchers can design experiments to remember each individual's responses, in order to customize the items or questions they are shown on subsequent rounds (Schembri et al., 2016b). Designs featuring real-time interactive group tasks (Hawkins, 2014) allow researchers to analyze collective learning (Zhu et al., 2014), or other large-scale social interactions, such as in complex economic games (Rand, 2012). For some designs, gamification elements such as leaderboards can be used to encourage competition among other participants (Melenhorst, Novak, Micheel, Larson, & Boeckle, 2015).

Attrition rates

Attrition rates in longitudinal studies can be a major concern. The available data on attrition rates on AMT is presented below, along with suggested measures to reduce attrition.

Attrition rates in offline studies

Figures for attrition rates in offline studies vary considerably, and can depend on a number of factors (Collins, Ellickson, Hays, & Mccaffrey, 2000; Ribisl et al., 1996). A meta-analysis of 85 school-based longitudinal studies (Hansen, Tobler & Graham, 1990) found an average retention rate of 75% after a year. Response rates are particularly high among studies with student participants, while researchers studying a more diverse population pool may report lower response rates. Collins et al. (2000) studied adolescents from 8 school districts representing diverse communities, and reported a 61% response rate after a four-month period. Commercial research panels, which allow researchers access to a wider and more varied participant pool, suffer from high rates of participant loss, with estimated response rates of 50% after a two-month period, and 15% after 13 months (Daly & Nataraajan, 2015).

Attrition rates on AMT

Among the longitudinal studies conducted on AMT, response rates are lower than reported in many traditional studies. Researchers have achieved response rates of 80% and 61% after a one-week period (Shapiro et al., 2013; Carr, 2014), 60-69% after a three-week period (Buhrmester et al., 2011; Holden et al., 2013); 75%, 56%, 38% and 47% after a two-, four-, eight-, and thirteen-month period respectively (Daly & Nataraajan, 2015), and 44% after a twelve-month period (Chandler et al., 2014).

Non-response bias on AMT

High attrition rates may be concerning in part because of the potential of a non-response bias, ie, the possibility that the set of participants who drop out of the study differ significantly from those who do not. Daly & Nataraajan (2015) explored this issue in a series of longitudinal studies on AMT. They found that participants who dropped out did not differ significantly from the overall sample on a number of factors. However, participants who completed a study tended to be 1-3 years older than those who did not. This bias will not be a concern for many studies, but should be taken into consideration. High attrition rates can be overcome by including a larger set of participants in the initial study (St Pierre, 1980). This requirement may well result in significant logistical issues in an offline study. However, AMT excels in giving researchers fast, cheap and easy access to a massive population of potential participants. Retention rates may also be improved by increasing the payment offered to complete a study (Collins et al., 2000). Given the relative lack of a non-response bias, and the ease of obtaining a larger pool of participants in the initial sample, lower response rates may not be a concern for many researchers on AMT.

Longi: reducing attrition rates

Longi includes a number of features designed to boost retention rates. Chandler et al. (2014) note that response rates were significantly higher (59%) among participants who had completed at least one posted task, or *Human Intelligence Task* (HIT), prior to the initial survey, as compared to those who had not (44%). These rates increased with the total numbers of HITs completed; the top 10% most productive participants had a response rate of 75%. This suggests that researchers can achieve high response rates by accepting only participants who have previously completed a high number of HITs. Longi allows researchers to filter participants based on the number of HITs they have previously completed, as well as the percentage of those HITs which were approved.

Attrition rates can be significantly reduced when researchers make repeated efforts to contact individuals over a period of time. Cotter, Burke, Stouthamer-Loeber and Loeber (2005) note that, when conducting their offline longitudinal study, 12% of participants required 20 or more contact attempts in order to complete the study. When conducting offline longitudinal studies, such contact attempts can be highly costly and time-consuming (Ribisl et al., 1996). Longi automates this process, allowing researchers to make regular contact attempts indefinitely. The software is able to reach participants even when they change the email address used for AMT.

Researchers can opt to pay a bonus to participants who have completed all rounds of the study. This can serve as an incentive to complete the study, particularly if researchers effectively communicate the number of rounds to complete, and the size of the reward.

Longi: An Overview

Longi is an open-source system which handles many of the technical challenges involved in running longitudinal studies on Amazon Mechanical Turk. A chief benefit of this system is its simplicity. Researchers can use Longi without any kind of programming background. Once their Amazon Web Services (AWS) account is set up, researchers who use Longi do not need to use the AMT website for any purpose other than adding funds.

Features and capabilities

Longi includes a series of Python-based scripts which handle basic “back-end” tasks common to AMT experiments, such as posting HITs, paying participants and automatically handling bonuses. In addition, researchers conducting longitudinal studies on AMT have a number of requirements: multiple HITs must be posted at time intervals specified by the researcher, these HITs must not be made available to all participants, initial rounds must be available only to those who have not completed previous experiments, subsequent rounds must be available only to those who have completed the immediately prior round, participants must be closely monitored, those who are eligible to complete subsequent rounds must be contacted periodically. Longi automates all of these tasks.

Longi creates a new HIT for each round of the experiment, according to the schedule specified by the researcher. *Windows scheduled tasks* are used to automate this procedure. If required, these can be viewed or modified in the Task Scheduler window independent of Longi. Researchers can choose to let Longi automatically approve completed HITs. If researchers need to verify participants’ responses before accepting the HIT, they can opt to do this manually instead. During every round of the experiment, a new custom qualification is created. This is automatically assigned to all participants who complete the round, and is then added as a requirement for subsequent rounds. Thus, any given round is available only to participants who have completed all previous rounds. While a round is ongoing, Longi keeps track of which participants have completed the current round. Reminders are sent out, at a set schedule, to eligible participants who have not yet done so; this continues until all participants complete the current round, or until a new round is posted.

Longi does not handle any of the "front-end" issues involved in creating a web-based experiment. The script assumes that a suitable experiment has already been designed, and simply takes in a URL as input. Longi can be used in conjunction with external platforms designed to build web-based experiments, such as Qualtrics (www.qualtrics.com), jsPsych (de Leeuw, 2014), domain-specific software such as Experigen (Becker & Levine, 2010) for language-learning experiments, or customized code designed by the individual researcher. Similarly, Longi does not provide a solution for data storage, but is able to interface with solutions such as psiTurk (Gureckis et al., 2015) or Submiterator (Lassiter, 2014).

Currently, Longi is able to run only one longitudinal study at a time. Further, only one round can be active at any given point. These restrictions will be lifted in future versions.

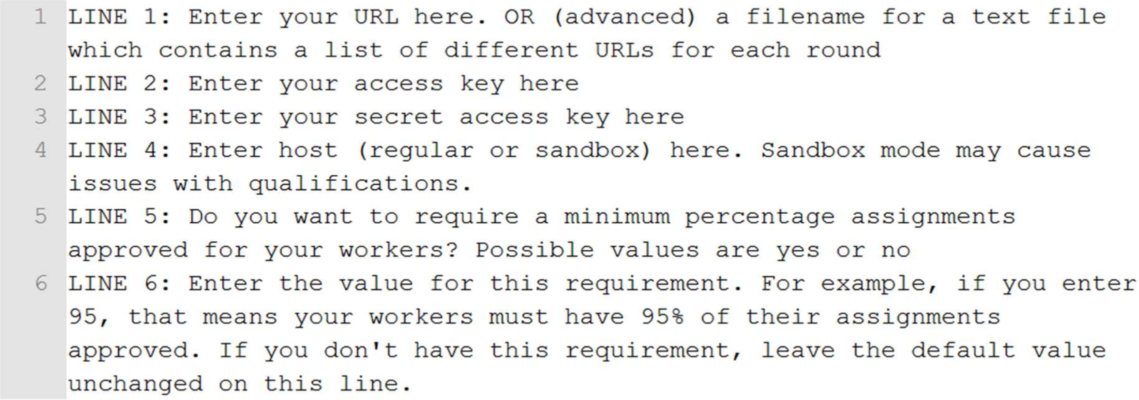
Technical requirements

Longi is designed to run on Windows systems that satisfy the following requirements:

1. Administrator privileges are required in order to create scheduled tasks programmatically.
2. A working installation of Python 2.7. This is the most recent version supported by boto (see below). Newer versions may work, but are not supported.
3. A working installation of boto, which is a Python interface to the AWS API. The readme files included in the package detail how to install boto, and how to check that it is running correctly.
4. Researchers must sign up for an Amazon Web Services account, and retrieve their access key and secret access key from https://console.aws.amazon.com

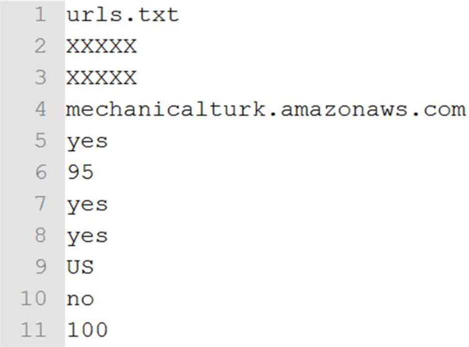
Setup

Python 2.7 and boto must be installed before setting up Longi. The readme instructions included with Longi explain how to check that boto is running correctly, and is able to interface with your AWS account. Before running Longi, researchers must fill out a configuration text file. Two example text files – for simple and advanced configuration options – are included with the Longi download. Each line of the text file must contain specific information in a set format. For example, the website URL must be entered on line 1, including the https:// prefix required by Amazon. An included text file explains what information should be entered on each line, and the possible formats. Once the configuration file has been filled out, researchers must navigate to their python installation directory at the command prompt and type python longi_scheduler.py. The longitudinal study is then able to proceed automatically, and no further action on the part of the researcher is required.



```
LINE 1: Enter your URL here. OR (advanced) a filename for a text file
which contains a list of different URLs for each round
LINE 2: Enter your access key here
LINE 3: Enter your secret access key here
LINE 4: Enter host (regular or sandbox) here. Sandbox mode may cause
issues with qualifications.
LINE 5: Do you want to require a minimum percentage assignments
approved for your workers? Possible values are yes or no
LINE 6: Enter the value for this requirement. For example, if you enter
95, that means your workers must have 95% of their assignments
approved. If you don't have this requirement, leave the default value
unchanged on this line.
```

Fig. 1 Extract from the included file which explains how to fill out the configuration text file



```
urls.txt
XXXXX
XXXXX
mechanicalturk.amazonaws.com
yes
95
yes
yes
US
no
100
```

Fig. 2 Extract from the included advanced configuration text file. The access key and secret access key are not shown in this file, due to security concerns.

Configuration options

Experiments can be configured in a number of ways. One set of options allows researchers to restrict the participants who are able to accept the HIT. Potential participants can be restricted by location, number and percentage of HITs approved. Researchers can also choose to block participants who have completed their HITs in the past, ensuring a fresh pool of participants for every experiment. When this option is selected, Longi creates a custom qualification, and assigns it to participants who complete any of its HITs. As long as this option is selected, participants who have received this qualification will not be able to complete HITs created by Longi. Note that this process happens automatically only with HITs that are created and processed by Longi. Researchers who have previously used qualifications to block participants can use an existing qualification for this purpose. This ensures that participants who have been blocked outside of Longi will not be able to accept HITs.

Another set of options involves the timing and frequency at which new rounds and reminders are posted. Researchers must specify how often new rounds should be posted. Rounds can be posted in increments of minutes, as well as hourly, daily, weekly or monthly; for example, every 2 weeks, or every month. A start and end date must also be provided. Longi allows researchers to write a custom message to participants. When a new round is posted, participants who have completed the previous round are automatically notified by email. Researchers can additionally choose to send reminders to participants at regular intervals; for example, every day. Longi will only send reminders to participants who have not yet completed the current round.

Longi is able to approve HITs automatically. If auto-approval is turned on, HITs are approved every 30 minutes. Additionally, researchers can choose to pay participants a bonus for completing all rounds of the experiment. This can serve as an incentive to avoid low response rates. If this option is selected, researchers should inform participants about the total number of rounds and the bonus amount. This is not handled automatically, but can be included in the HIT title and description, as well as in emails. Longi additionally handles the standard set of options involved with creating a new HIT. Researchers can specify the title, description and keywords for their HIT, and set the payment amount, total number of participants, time limit and expiry date.

Finally, researchers must specify a deletion date for their experiment. At this point, Longi will delete all data pertaining to the current experiment, including all data tracking participants and HITs. Custom qualifications created for the experiment are also deleted. This avoids one of the pitfalls of using qualifications: participants are notified when qualifications are revoked. In a longitudinal experiment with multiple rounds, and one qualification per round, multiple unnecessary notifications can cause a great deal of annoyance for participants. When temporary qualifications are deleted rather than revoked, no notifications are generated. As a consequence, however, continuing an experiment past the deletion date becomes rather difficult, and must be done manually rather than through Longi. If unsure of the length of an experiment, researchers should pick a deletion date in the far future.

Advanced options

The basic configuration allows for a single setup which remains constant throughout every round of the experiment. Advanced options allow the researcher to specify different values at different points in the experiment. For example, researchers may want to send a different email message for each round, rather than using a standard message throughout the experiment. Similarly, HITs can contain a unique title and description for each round. Advanced options also allow for highly individualized posting schedules for both HITs and reminders, making it possible to post new HITs at irregular intervals. The included explanation file contains further details on the advanced options which are available.

Usage scenario

Longi has been successfully used to manage a set of 2 language learning experiments conducted on AMT over the course of 4 consecutive days (Schembri et al., 2016a; Schembri et al., 2016b). In each experiment, participants were taught an artificial stress system based on Cairene Arabic over 4 20-30-minute sessions. Participants were paid $1.50 for completing each round and a further $1.50 bonus for completing all four rounds. Information about the completion bonus was posted prominently on all HIT titles and email subjects sent to participants; feedback and correspondence from participants indicated that they were aware of the bonus, and that it served as an incentive for completion. Participants were required to be US residents with a prior HIT approval rate of at least 95%. HITs were posted every 24 hours and were available to participants for a 24 hour period before expiring. Participants were asked to respond only if they had had a full night's sleep between rounds. Reminder emails were sent out every 4-5 hours on an irregular schedule. In the first experiment, 82 participants finished all four rounds of the experiment; data from a further 9 participants was excluded. In the second experiment, 83 participants finished all four rounds of the experiment; data from a further 11 participants was excluded. Attrition rates were fairly high in the first experiment due to technical issues with the experiment website: 57% of participants completed all 4 rounds in the first experiment, while the completion rate for the second experiment was 70%.

Conclusion

AMT is an excellent platform for longitudinal studies, and as such it should be considered as an option both for researchers conducting more traditional longitudinal studies, as well as those who may not otherwise consider a longitudinal design. Our hope is that Longi will enable more of these designs by simplifying the significant technical challenges involved.

Appendix

API functions

Longi uses boto, a Python interface to Amazon Web Services, to access the AWS API. The script utilizes a number of native functions provided by the API, as detailed in the table below.

API Function	**Description**	**Parameters used by Longi**
MTurkConnection	Connect to an AWS account	access_key, secret_access_key, host
create_hit	Create a new HIT with the specified options	lifetime, max_assignments, keywords, reward, duration, approval delay, title, description, qualifications, response_groups
create_qualification_type	Create a new qualification	name, description, status
assign_qualification	Assign a qualification to a worker ID	qualification_type_id, worker_id, send_notification
get_assignments	Return all completed assignments associated with a HIT ID	hit_id, page_number
approve_assignment	Approve an assignment	assignment_id
notify_workers	Send an email to an address associated with a worker ID	worker_id, subject, message
grant_bonus	Grant a bonus	worker_id, assignment_id, payment, message
dispose_qualification_type	Delete a qualification	qualification_id

Table 1 List of API functions used by Longi

In principle, Longi could be adapted for any crowdfunding platform which provides an API with similar functionality.

Program structure

Longi consists of 5 individual scripts. As it runs, it also creates a number of batch files and text files to store data and simplify task execution.

While an experiment is ongoing, 3 scripts handle all the main functions. The hit creation script creates a HIT based on the specified options, and saves the HIT ID for later use. It also creates and assigns qualifications where necessary. The approval script approves completed HITs. For each round of the experiment, this script compiles and maintains a list of worker IDs associated with an approved HIT. The reminder script is triggered only for second and subsequent rounds. It compares the current round's list of worker IDs, as compiled by the approval script, with the previous round's list. Worker IDs which appear in the previous round's list, but not in the current round's list, are sent a reminder through the notify_workers function. Each of these 3 main scripts is associated with a separate *Windows Scheduled Task*, and so can run on an individualized schedule, based on the options entered in the configuration file.

The remaining scripts handle initialization and finalization of the experiment. The initialization script reads in the options specified by the user, and creates batch files and Windows Scheduled Tasks based on these options. These scheduled tasks call the remaining scripts where necessary; the user interacts only with the initialization script. The finalization script deletes the batch files, text files and qualifications used, and allows researchers to begin a new experiment.

References

Albertyna Paciorek, J. N. W. (2015). Semantic Generalization in Implicit Language Learning. *Journal of Experimental Psychology Learning Memory and Cognition*.

Bauer, K. W. (2004). Conducting longitudinal studies. *New Directions for Institutional Research*, *2004*(121), 75–90.

Becker, M., & Levine, J. (2010). Experigen: an online experiment platform. *Available (April 2013) at https://github. com/tlozoot/experigen.*

Boynton, M. H., & Richman, L. S. (2014). An online daily diary study of alcohol use using Amazon's Mechanical Turk. *Drug and Alcohol Review*, *33*(4), 456–461.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.

Carr, A. (2014). An Exploration of Mechanical Turk as a Feasible Recruitment Platform for Cancer Survivors.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156–2160.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130.

Chandler, J., & Shapiro, D. (2016). Conducting Clinical Research Using Crowdsourced Convenience Samples. *Annual Review of Clinical Psychology*, *12*(1).

Collins, R. L., Ellickson, P. L., Hays, R. D., & Mccaffrey, D. F. (2000). Effects of Incentive Size and Timing on Response Rates to a Follow-Up Wave of a Longitudinal Mailed Survey. *Evaluation Review*, *24*(4), 347–363.

Cotter, R. B., Burke, J. D., Stouthamer-Loeber, M., & Loeber, R. (2005). Contacting participants for follow-up: how much effort is required to retain participants in longitudinal studies? *Evaluation and Program Planning*, *28*(1), 15–21.

Cress, U., Held, C., & Kimmerle, J. (2013). The collective knowledge of social tags: Direct and indirect influences on navigation, learning, and information processing. *Computers & Education*, *60*(1), 59–73.

Culbertson, J., & Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences*, *111*(16), 5842–5847.

Culbertson, J., & Wilson, C. (2013). *Artificial grammar learning of shape-based noun classification*. Manuscript.

Daly, T. M., & Nataraajan, R. (2015). Swapping bricks for clicks: Crowdsourcing longitudinal data on Amazon Turk. *Journal of Business Research*.

de Leeuw, J. R. (2014). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12.

Demment, M. (2015). Using Amazon's Mechanical Turk as a tool for a global survey: Lessons learned from a large-scale implementation. Presented at the 2015 APHA Annual Meeting & Expo (Oct. 31 - Nov. 4, 2015), APHA.

Enochson, K. (2015). Adaptation as statistical learning: An individual differences study.

Fenn, K. M., Margoliash, D., & Nusbaum, H. C. (2013). Sleep restores loss of generalized but not rote learning of synthetic speech. *Cognition*, *128*(3), 280–286.

Foreman, N., Stanton, D., Wilson, P., & Duffy, H. (2003). Spatial Knowledge of a Real School Environment Acquired From Virtual or Physical Models by Able-Bodied Children and Children With Physical Disabilities. *Journal of Experimental Psychology. Applied*, *9*(2), 67–74.

Forget, A., Chiasson, S., & Biddle, R. (2012). Supporting learning of an unfamiliar authentication scheme. *AACE E-Learn, E-Learn*.

Fort, M., Martin, A., & Peperkamp, S. (2015). Consonants are More Important than Vowels in the Bouba-kiki Effect. *Language and Speech*, *58*(2), 247–266.

Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*, 80–96.

Gallagher, G. (2013). Learning the identity effect as an artificial language: bias and generalisation. *Phonology*, *30*(02), 253–295.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making*, *26*(3), 213–224.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., … Chan, P. (2015). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, 1–14.

Hansen, W. B., Tobler, N. S., & Graham, J. W. (1990). Attrition in substance abuse prevention research a meta-analysis of 85 longitudinally followed cohorts. *Evaluation Review*, *14*(6), 677–685.

Hareli, S., Kafetsios, K., & Hess, U. (2015). A cross-cultural study on emotion expression and the learning of social norms. *Frontiers in Psychology*, *6*. http://doi.org/10.3389/fpsyg.2015.01501

Hawkins, R. X. D. (2014). Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, *47*(4), 966–976.

Holden, C. J., Dennie, T., & Hicks, A. D. (2013). Assessing the reliability of the M5-120 on Amazon's mechanical Turk. *Computers in Human Behavior*, *29*(4), 1749–1754.

Horowitz, A., & Frank, M. C. (2012). Learning from speaker word choice by assuming adjectives are informative. In *Proceedings of the 34th annual conference of the cognitive science society*. Citeseer.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*(2), 151–195.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790.

Lanaj, K., Johnson, R. E., & Barnes, C. M. (2014). Beginning the workday yet already depleted? Consequences of late-night smartphone use and sleep. *Organizational Behavior and Human Decision Processes*, *124*(1), 11–23.

Lassiter, D. (2014) Submiterator. Retrieved from http://github.com/danlassiter/Submiterator

Linzen, T., & Gallagher, G. (2014). The Timecourse of Generalization in Phonotactic Learning. *Proceedings of the Annual Meetings on Phonology*, *1*(1).

Lynn, B. M.-D. (2014). *Shared Sense of Purpose and Well-Being among Veterans and Non-Veterans*. North Carolina State University.

Maass, J. K., Pavlik Jr, P. I., & Hua, H. (2015, June). How Spacing and Variable Retrieval Practice Affect the Learning of Statistics Concepts. In *Artificial Intelligence in Education* (pp. 247-256). Springer International Publishing.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23.

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(1), 128.

Melenhorst, M., Novak, J., Micheel, I., Larson, M., & Boeckle, M. (2015). Bridging the Utilitarian-Hedonic Divide in Crowdsourcing Applications. In *Proceedings of the Fourth International Workshop on Crowdsourcing for Multimedia* (pp. 9–14). New York, NY, USA: ACM.

Odierna, D. H., & Schmidt, L. A. (2009). The Effects of Failing to Include Hard-to-Reach Respondents in Longitudinal Surveys. *American Journal of Public Health*, *99*(8), 1515–1521.

Parent, J., McKee, L. G., Rough, J. N., & Forehand, R. (2016). The association of parent mindfulness with parenting and youth psychopathology across three developmental stages. *Journal of Abnormal Child Psychology*, *44*(1), 191–202.

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, *299*, 172–179.

Raybourn, E. M. (2007). Applying simulation experience design methods to creating serious game-based adaptive training systems. *Interacting with Computers*, *19*(2), 206–214.

Ribisl, K. M., Walton, M. A., Mowbray, C. T., Luke, D. A., Davidson II, W. S., & Bootsmiller, B. J. (1996). Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: Review and recommendations. *Evaluation and Program Planning*, *19*(1), 1–25.

Rottman, B. M. (2014). Information Search in an Autocorrelated Causal Learning Environment. In *Proceedings of the 36th annual conference of the cognitive science society, Cognitive Science Society, Austin, TX.*

Schembri, T., Johnson, M., & Demuth, K. (2016a). Online learning of Cairene Arabic word stress patterns over time. In preparation.

Schembri, T., Johnson, M., & Demuth, K. (2016b). Is Less Really More? Answers from L2 Stress Acquisition. In preparation.

Schleider, J. L., & Weisz, J. R. (2015). Using Mechanical Turk to Study Family Processes and Youth Mental Health: A Test of Feasibility. *Journal of Child and Family Studies*, *24*(11), 3235–3246.

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*. 1:213–20

Simcox, T., & Fiez, J. A. (2014). Collecting response times using Amazon Mechanical Turk and Adobe Flash. *Behavior Research Methods*, *46*(1), 95–111.

Smith, N. A., Sabat, I. E., Martinez, L. R., Weaver, K., & Xu, S. (2015). A Convenient Solution: Using MTurk To Sample From Hard-To-Reach Populations. *Industrial and Organizational Psychology*, *8*(02), 220–228.

Stacey, P. C., Raine, C. H., O'Donoghue, G. M., Tapper, L., Twomey, T., & Summerfield, A. Q. (2010). Effectiveness of computer-based auditory training for adult users of cochlear implants. *International Journal of Audiology*, *49*(5), 347–356. http://doi.org/10.3109/14992020903397838

Standing, L., Conezio, J., & Haber, R. N. (1970). Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. *Psychonomic Science*, *19*(2), 73–74.

St.Pierre, R. G. (1980). Planning Longitudinal Field Studies Considerations in Determining Sample Size. *Evaluation Review*, *4*(3), 405–415.

Tenenbaum, R. Z., Byrne, C. J., & Dahling, J. J. (2014). Interactive Effects of Physical Disability Severity and Age of Disability Onset on RIASEC Self-Efficacies. *Journal of Career Assessment*, *22*(2), 274–289.

Tily, H., Frank, M. C., & Jaeger, T. F. (2011). The learnability of constructed languages reflects typological patterns. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1364–1369).

van Weel, C. (2005). Longitudinal research and data collection in primary care. *The Annals of Family Medicine*, *3* (suppl 1), S46–S51.

Vaughn, A. A., Cronan, S. B., & Beavers, A. J. (2015). Resource Effects on In-Group Boundary Formation With Regard to Sexual Identity. *Social Psychological and Personality Science*, *6*(3), 292–299.

Vogel, A., Emilsson, A. G., Frank, M. C., Jurafsky, D., & Potts, C. (2014). Learning to reason pragmatically with cognitive limitations. In *Proceedings of the 36th annual meeting of the cognitive science society* (pp. 3055–3060). Cognitive Science Society Austin, TX.

Zhu, H., Dow, S. P., Kraut, R. E., & Kittur, A. (2014). Reviewing versus doing: Learning and performance in crowd assessment. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 1445–1455). ACM.

Chapter 3: Online acquisition of Cairene Arabic word stress patterns over time.

ABSTRACT

Second language (L2) acquisition often poses significant challenges for adult learners; native-like attainment is commonly held to be unachievable. The L2 acquisition of main word stress is particularly challenging. Most previous research has focussed on the acquisition of word stress in English and other European languages; we therefore have a limited understanding of how other stress systems are acquired. Cairene Arabic has a complex stress system which has been extensively studied. In the current study, American English speakers (N = 73) were taught the Cairene Arabic stress system over a 4-day period. Participants had no prior experience with the target L2 or any Semitic languages. Artificial language learning methodology was combined with input derived from a corpus of Colloquial Cairene Arabic. The experiment aimed to quantify the factors affecting participants' acquisition. Evidence for differential rates of acquisition between two sets of constraint rankings was demonstrated in words with final stress. Patterns of acquisition similar to prior studies were observed. Participants' responses to words with heavy syllables showed evidence of L1 transfer. Evidence of overgeneralization was seen in participants' avoidance of initial stress, leading to distributional patterns seen in neither the L1 nor the target language.

Adult learners often experience difficulty in attaining fluency in a second language. The acquisition of main word stress is one area of particular challenge. This is significant, as prosody has been shown to have a greater effect on intelligibility than segmental factors (Munro & Derwing, 1999). Word stress interacts with a number of phonological constraints, such as syllable weight, foot construction, rhythm, and phonological processes such as syncope and epenthesis. Because of this, research into the acquisition of word stress provides a critical window into learners' underlying knowledge of prosodic structure.

Infants are sensitive to the prosodic and suprasegmental properties of their native language from an early age. At 7.5 months, infants can use prosodic information to segment words from fluent speech at the onsets of stressed syllables (Jusczyk, Houston & Newsome, 1999). When stress cues are designed to conflict with transitional probabilities, a statistical cue to word segmentation, 8-month-old infants rely on the prosodic, rather than statistical information (Johnson & Jusczyk, 2001). At 9 months, American infants show a preference for the predominant strong-weak (SW) stress pattern of English (Jusczyk, Cutler & Redanz, 1993). This suggests that, at least for languages like English, sensitivity to some of the prosodic cues to stress and syllable prominence develop within the first year of life. Infants become attuned to the rhythmic patterns of their native language very early on, thus making it more difficult to process non-native patterns. Infants' early acquisition of stress, in the first year of life, may account for the difficulty adult learners have in acquiring stress in a second language.

Much of the literature studying the L2 acquisition of word stress is concerned with the acquisition of word stress in English (Altmann, 2006; Archibald, 1997; Guion, Harada, & Clark, 2004; Guion, 2005; Wayland, Landfair, Li, & Guion, 2006; Pater, 1997; Tremblay & Owens, 2010). Research in this area has largely concentrated on two main areas: transfer from the L1 (Anani, 1989; Archibald, 1993; Baptista, 1989; Erdmann, 1973; Mairs, 1989; Pater, 1997), and factors affecting the acquisition of L2 stress systems (Altmann, 2006; Guion et al., 2004; Guion, 2005; Peperkamp & Dupoux, 2002; Peperkamp, Vendelin & Dupoux, 2010; Tremblay & Owens, 2010). Early research asked participants to read out real words in the target language from lists, sentences or paragraphs. The main findings were that, while certain patterns of stress placement can be easily attributable to L1 transfer (Anani, 1989; Archibald, 1993; Baptista, 1989; Mairs, 1989), learners sometimes produced interlanguage forms which are unlike either the native or target language

(Archibald, 1993; Erdmann, 1973; Pater, 1997). This early research was criticized due to its use of real words, which made it difficult to tell whether learners had simply memorized the lexical stress patterns or not. As a result, later research largely used nonce words as stimuli.

Subsequent research has also focussed on quantifying the factors affecting the successful perception and production of word stress. One major research area has concerned the perception of stress in nonce words in participants with a wide range of language backgrounds. Native speakers of languages with predictable stress placement, such as French, performed poorly on stress perception in nonce words (Altmann, 2006; Dupoux, Peperkamp & Sebastien-Galles, 2001; Dupoux & Peperkamp, 2002; Peperkamp & Dupoux, 2002; Peperkamp, Vendelin & Dupoux, 2010). Native speakers of languages with non-predictable stress, such as English or Spanish, or without stress, such as Thai or Chinese (Altmann, 2006), performed at ceiling. In another set of studies, groups of early and late Spanish advanced learners of English (Guion et al., 2004) and Korean advanced learners of English (Guion, 2005) were tested on the perception and production of word stress. Perception and production were found to be independently affected by a number of factors, such as phonological similarity to known words in either the native or target language, lexical class, and the presence or absence of long vowels and coda consonants. These factors influenced early and late learners differently. For example, Spanish late learners ignored the presence of long vowels and coda consonants in both production and perception, while early learners patterned more closely with native speaker controls.

In the current study, adult native English participants are asked to complete a learning task, acquiring aspects of the colloquial Cairene Arabic stress system over the course of 4 days. Cairene Arabic has been extensively studied due to its complex stress system; however, there is little research on its acquisition. The following sections describe the English and Cairene Arabic word stress systems, detailing the similarities and differences between the two languages. Based on these comparisons, predictions are made about the effect of L1 transfer in the current experiment.

ENGLISH WORD STRESS

English primary word stress is not predictable from purely phonological structure (Peperkamp & Dupoux, 2002). This can most clearly be demonstrated through the existence of minimal pairs such as *con*tent and con*tent*. However, although English stress placement is variable, a number of distributional regularities can be observed that might help a learner of English. For example, syllables with long vowels are more likely to attract primary stress. An analysis of the CELEX lexical database reveals that long vowels are roughly twice as likely (60%) to be stressed compared to short vowels (35%) (Guion et al., 2003). Duration is an important cue for stress in English: unstressed vowels are reduced, in duration as well as quality. As a result, long vowels may be especially salient for English speakers. This may explain why long vowels attract stress more often than short vowels. Syllables with a coda consonant are more likely to attract stress than open syllables. Initial stress is the most frequent pattern, which accounts for 57% of polysyllabic English content words (Cutler & Carter, 1987). Stress varies systematically with grammatical class (Burzio, 1994; Chomsky & Halle, 1968; Hayes, 1982; Hayes, 1995). Nouns are more likely than average to receive initial stress. While 78% of all bisyllabic words receive initial stress (Clopper, 2002), 92% of bisyllabic nouns are stressed initially (Sereno, 1986). Nouns with a heavy penultimate syllable tend to receive penultimate stress (Burzio, 1994; Hayes, 1995).

Nonce words can be used to test speakers' productive knowledge of stress. Given the English stress patterns outlined above, English speakers are expected to be sensitive to the different distributional properties of nouns and verbs, and this is the case: they are significantly more likely to select initial stress for nonce words in a noun frame rather than a verb frame (Baker & Smith, 1976; Davis & Kelly, 1997; Guion et al., 2003). When nonce words were presented in a noun frame, participants rarely assigned main stress to the final syllable (Domahs et al., 2014; Pater, 1997). Syllable weight has little effect on final stress avoidance, even where a final syllable is superheavy (CVCC or CVVC): participants assigned final stress for only around 20% of final CVCC syllables (Domahs et al., 2014; Guion et al., 2003). Syllable weight has a greater effect in non-final position: heavy (CVC or CVV) penultimate syllables were stressed almost categorically (Domahs et al., 2014; Pater, 1997). Where there is no heavy penultimate syllable, participants most commonly assigned initial stress for 2 and 3 syllable words (Pater, 1997; Guion et al., 2003; Domahs et al., 2014).

CAIRENE ARABIC WORD STRESS

Colloquial Cairene Arabic is a quantity-sensitive language with a three-way distinction between light (CV), heavy (CVC, CVV) and superheavy (CVCC, CVVC) syllables. In some analyses, superheavy syllables are treated as heavy syllables with extrametrical final consonants (Hayes, 1995). Primary stress can appear in final, penultimate or antepenultimate position. Secondary stress is not realized in the surface form (Halle & Vergnaud, 1987; Crowhurst, 1996). The location of stress is entirely predictable by metrical structure, and can be determined if the sequence of light (L), heavy (H) and superheavy (S) syllables is known. For example, a word with 3 light syllables (LLL) is always assigned antepenultimate stress, while a word with two light syllables and a final superheavy (LLS) is always assigned final stress. The surface observations describing the location of colloquial Cairene Arabic primary stress are summarized below (Broselow, 1976; Broselow, 1979; McCarthy, 1979; Hayes, 1995). The generalization in (4) is a simplified version of the commonly cited observation for classical Cairene Arabic; it holds true for the words in the corpus used for this experiment.

1) heavy final syllables are never stressed

'fi.him	*‘He understood’*
mu.'dar.ris	*‘Teacher’*

2) final superheavy or CVV syllables are always stressed

ka.'tabt	*‘I wrote’*
sa.ka.'kiin	*‘Knives’*
ra.'maa	*‘He threw him’*

3) otherwise, heavy penultimate syllables are stressed

ka.'tib.lik	*‘He wrote to you’*
'bee.tak	*‘Your house’*

4) where the above do not apply:

a) if the antepenultimate is heavy, stress the penultimate

jik.'ti.bu	*‘They write’*
mar.'ta.ba	*‘Mattress’*

b) if the antepenultimate is light, stress the antepenultimate

'ga.sa.di	*‘Physically’*
'da.ra.sit	*‘She studied’*

These observations can be summarized as follows. Final superheavy or CVV syllables are always stressed. For all other cases, the antepenultimate is stressed if it is light; otherwise, the penultimate is stressed. The patterns in (4) are unusual cross-linguistically. Typically, in weight-sensitive languages, heavy syllables are stressed over light syllables. However, light antepenultimate syllables attract stress in Cairene Arabic, even though heavy antepenultimate syllables do not. These patterns can be explained through footing constraints, which are described in greater detail below. However, on the surface, this appears to be an exception to the regular weight system. Because of this, the Cairene Arabic stress system has been studied extensively. It is possible that this apparent exception may make these patterns more difficult to acquire in both L1 and L2 acquisition. In addition, corpus data (Kilany, Hanaa et al., 1997) demonstrates that the patterns in (4) are significantly less frequent than the other patterns; the stimuli in the current experiment reflect these natural frequencies. The low frequency of these patterns is likely to further contribute towards difficulty in acquisition. Additionally, English native speakers may find it difficult to acquire the patterns in (2), in which final superheavy or CVV syllables are always stressed, due to the pattern of final stress avoidance in English nouns. Novel word experiments demonstrate that native English participants avoid stressing a final syllable even when it is superheavy (Domahs et al., 2014; Guion et al., 2003). Because of this tendency, it may be difficult to learn a system in which final superheavy syllables are categorically stressed.

CAIRENE ARABIC WORD STRESS: OPTIMALITY THEORY ANALYSIS

An Optimality Theory (OT) analysis of Cairene Arabic stress was carried out to organize the stimuli into constraint groups. Cairene Arabic stress realization requires the construction of trochaic feet from left-to-right. This can be achieved with the interacting constraints ALIGN-L and TROCHAIC. The constraint FTBIN-μ ensures that feet contain exactly two moras. It is violated when feet do not consist of either two light syllables, or one heavy syllable (Hayes, 1995). The definitions of these constraints are given below. The tableau in (4) demonstrates this effect, and shows how light antepenultimate syllables receive stress due to the interaction of footing constraints. In the analysis below, parsed feet are enclosed with parenthesis ‘( )’, syllable boundaries are marked with a period ‘ . ’, and stressed syllables are marked with the IPA symbol ‘ ˈ ’ preceding the stressed syllable.

(1) **Align-L**

Every Prosodic Word (ω) must begin with a foot. (Kager, 2001; Kager, 2005)

(2) **Trochaic**

Construct trochaic feet. (Tesar, 1997)

(3) **FtBin-μ**

Feet must be binary under a moraic (μ) analysis. (Hewitt, 1994)

(4) **Align-L, Trochaic**

CV.CV.CV	**Align-L**	**Trochaic**
☞ (ˈCV.CV).CV		
(CV.ˈCV).CV		!*
CV.(ˈCV.CV)	!*	
CV.(CV.ˈCV)	*	*

The constraint NONFIN is violated when the final syllable is contained in the head foot; thus, it bans main stress on the final syllable. WSP states that heavy syllables must be stressed. Final heavy syllables remain unstressed due to the ranking NONFIN >> WSP, as demonstrated in the tableau below.

(5) **NONFIN**

No prosodic head of the Prosodic Head (ω) is final in ω. (Prince & Smolensky, 2002).

(6) **WEIGHT-TO-STRESS PRINCIPLE (WSP)**

No unstressed bimoraic syllables. (Prince, 1990)

(7) **NONFIN >> WSP**

CV.CVC	**NONFIN**	**WSP**
☞ ('CV).CVC		*
(CV).('CVC)	!*	

The constraint WSPμμμ states that superheavy syllables must be stressed. The ad-hoc constraint WSP-CVV states that CVV syllables must be stressed. The rankings WSPμμμ >> NONFIN and WSP-CVV >> NONFIN are illustrated below:

(8) **WSPμμμ**

No unstressed trimoraic syllables. (Gouskova, 2003).

(9) **WSP-CVV**

No unstressed CVV syllables.

(10) **WSPμμμ >> NonFin**

CVC.CVCC	**WSPμμμ**	**NonFin**
☞ (CVC).('CVCC)		*
('CVC).(CVCC)	!*	

(11) **WSP-CVV >> NonFin**

CVC.CVCC	**WSP-CVV**	**NonFin**
☞ (CVC).('CVV)		*
('CVC).(CVV)	!*	

The constraint Endrule-R states that main stress must occur in the rightmost foot. When a word contains multiple non-final heavy syllables, it is EndRule-R which determines the location of main stress. The interaction between EndRule-R and NonFin ensures that penultimate heavy syllables are always stressed.

(12) **ENDRULE-R**

The head foot is not followed by another foot within the ω.

(McCarthy, 2003; Prince, 1983)

(13) **ENDRULE-R, NONFIN**

CVC.CVC.CVC	**ENDRULE-R**	**NONFIN**
☞ (CVC).('CVC).CVC		
(CVC).(CVC).('CVC)		!*
(CVC).('CVC).(CVC)	!*	
('CVC).(CVC).CVC	!*	

Heavy antepenultimate syllables remain unstressed due to the interaction of footing constraints with ENDRULE-R. A foot (H) is constructed around the CVC syllable, with additional feet constructed to the right. As a result, the heavy antepenultimate syllable is never contained in the rightmost foot, and therefore cannot bear main stress. The constraint LAPSE-FT ensures that two adjacent syllables cannot remain unfooted. Note that the top two candidates produce the same surface result, and the current constraint ranking does not distinguish between them.

(14) **LAPSE-FT**

Incur a violation for two adjacent unfooted syllables (de Lacy, 2002; Green & Kenstowicz, 1995)

(15) **LAPSE-FT >> ENDRULE-R, FTBIN, NONFIN**

CVC.CV.CV	**LAPSE-FT**	**ENDRULE-R**	**FTBIN**	**NONFIN**
☞ (CVC).('CV.CV)				*
☞ (CVC).('CV).CV			*	
('CVC).(CV.CV)		!*		!*
('CVC).CV.CV	!*			

The above analysis provides an explanation for the surface patterns observed for the Cairene Arabic stress system. This results in the full range of patterns observed: correct stress assignment is derived from one or two constraints in many of the examples above, or from the interplay of multiple constraints in (15). These groupings of constraints are used to organize the stimuli in the current experiment.

THE CURRENT STUDY

Research into the L2 acquisition of word stress faces some common methodological challenges. Participants in many studies may already have some degree of exposure to the L2. Researchers must therefore control for age of acquisition, length of exposure and proficiency in order to draw meaningful comparisons between participants (Altmann, 2006; Guion et al., 2004; Guion, 2005; Tremblay & Owens, 2010). The choice of stimuli can also be problematic. Some studies use real English words (Archibald, 1993; Mairs, 1989); however, this raises the possibility that participants are memorizing stress placement on a word-by-word basis, rather than applying a generalized stress rule (Pater, 1997). The use of nonce word stimuli introduces a different set of challenges. In a non-predictable stress language, researchers must determine how native speakers would stress a given nonce word. This is commonly done by running a parallel experiment with native speakers (Alvord, 2003; Guion, Clark, Harada, & Wayland, 2003; Guion, 2005; Domahs, Plag, & Carroll, 2014); however, this is costly and time-consuming. A nonce word must be accepted as a plausible word in the target language; therefore, infrequent phonotactic combinations are generally avoided. As a result, participants may be influenced by similar-sounding existing words (Altmann, 2006; Guion et al., 2004; Guion, 2005), or may interpret nonce words in different ways, thus affecting stress placement. For example, Altmann (2006) argues that some participants may interpret a nonce word such as tugumster as containing the derivational morpheme '-er', while others may interpret it as monomorphemic. Thus, while the use of real words is problematic, the construction of suitable nonce words also poses a number of challenges for researchers.

The artificial language learning (ALL) paradigm (Culbertson 2012; Reber, 1967) offers a different way of addressing these challenges. ALL experiments are designed to investigate learning biases that can shed light on typological patterns. As a result, the artificial languages used for these experiments are not intended to represent any real-world language, real or potential. Instead, simplified toy languages with minimal features are designed to answer a specific research question. Experiments typically consist of 1) a familiarization phase, where participants are passively exposed to the artificial language, 2) a testing phase, where participants are tested on their knowledge of stimuli presented in the familiarization phase, and 3) a generalization phase, where participants are tested on their ability to generalize to novel words. Novel words are previously unseen and unfamiliar to participants, but are structurally similar to words which have been already presented. The artificial language learning (ALL) paradigm has been widely used in research into the acquisition of phonology (Moreton, 2008; Wilson, 2006), morphology (Fedzechkina, Jaeger, & Newport, 2011; St Clair, Monaghan, & Ramscar, 2009) and syntax (Christiansen, 2000; Culbertson, Smolensky, & Legendre, 2012). However, few studies have focussed directly on teaching an artificial stress system. Guest, Dell, & Cole (2000) constructed two artificial stress systems by combining eight CV, CVV and CVC syllables in different permutations to create 3-7 syllable words. This study established that participants in an ALL experiment are able to successfully acquire a stress system: 62% of responses accurately reflected the stress patterns acquired during training. Carpenter (2005, 2010) constructed four artificial stress systems by combining 32 CV syllables in different permutations of three and four syllables. Participants stressed 61-70% of the novel words correctly, depending on condition. This study expanded on the methodology in Guest et al. (2000), formalizing a procedure for teaching an artificial stress system. These results, together with those in Guest et al. (2000), establish a baseline for successful acquisition of a stress system within the ALL paradigm.

The current study aims to combine some of the features of the ALL paradigm and L2 acquisition research, employing the controlled nature of the ALL paradigm with the complexity and messiness of real-world language acquisition. As in previous ALL research, participants were taught a language with which they have no prior experience. As a result, the length of exposure was identical for all participants, and there was no need to control for proficiency. In order to successfully acquire a stress system, participants must first accurately perceive stress. Speakers of languages with non-predictable stress outperform others in stress perception (Altmann, 2006; Peperkamp & Dupoux, 2002); hence, native English-speaking participants were ideal for this experiment. As in some L2 acquisition research, the stimuli were real-world lexical items. Because participants were selected to be unfamiliar with the target language, there were no issues with prior memorization of the target structure: stimuli were experienced as nonce words (Salsignac, 1998). Similarly, participants were unlikely to apply a strategy based on analogy to known words, as the target language is quite different from their native language. Stimuli were designed to capture the phonological complexity and relative frequencies of the target language to the greatest extent practical. The aim was to mimic real-world acquisition, rather than to narrowly target a specific set of phonetic or phonological structures. As in the ALL paradigm for phonological experiments, participants were not required to learn any semantic mapping, but were tested solely on their acquisition of the Cairene Arabic stress system. Successful acquisition was measured through participants' ability to generalize to novel items.

Earlier ALL experiments have succeeded in teaching participants a simple stress system with restricted input in terms of syllable structure, word length and phonemic inventory (Guest et al., 2000; Carpenter 2005; Carpenter, 2010). We predict that this success will carry over to the acquisition of a complex stress system with input that closely reflects the target language in terms of phonological complexity. However, given that the stress system is being taught over the course of multiple days, memory and retention issues may result in weaker performance than in previous studies. The current experiment aims to quantify the factors affecting participants' acquisition. Sets of OT constraints, and the rankings between them, were used to organize the stimuli into groups. These groupings were used to try to determine whether participants showed evidence of differential rates of acquisition for the various constraint groups. This would indicate that certain constraint rankings were easier to acquire than others. However, this proved difficult to determine for all sets of constraints.

A secondary analysis shed light on the factors affecting participants' acquisition by examining the differences between the English and Cairene Arabic stress system. L1 transfer plays an important role in L2 acquisition. In the current experiment, this effect can be quantified across a number of parameters. Participants may apply a purely L1 transfer strategy, treating the stimuli as English nonce words. If this holds true, the stress assignment strategies in the current study should mirror those seen with English nonce words presented in a noun frame. However, if participants begin to acquire the target language, their performance should demonstrate L1 transfer effects in some, but not all, contexts. Comparing the L1 and target language allows predictions to be made. In some instances, the L1 and target language pattern similarly. Heavy penultimate syllables are always stressed in Cairene Arabic, while in English, they are stressed almost categorically. As a result, words with a heavy penultimate syllable should be comparatively easy for participants to acquire. More significant are cases in which the L1 and target language diverge in their predictions. The stimuli in the current experiment are presented as nouns. English nouns overwhelmingly receive initial stress; final stress is uncommon. As discussed above, native English speaking participants reflect these distributional patterns in nonce word experiments. Cairene Arabic contains a sizable number of words with final stress. If participants apply an L1 strategy, they should perform poorly on finally stressed words, and perform well on words with initial stress. Conversely, if participants fully acquire the target structure, they should perform equally well on words with initial and final stress.

The English and Cairene Arabic stress systems also differ in their treatment of weight. In English, weight contributes to stress assignment in a probabilistic manner. In Cairene Arabic, the weight system is entirely deterministic: if the sequence of light, heavy and superheavy syllables in a word is known, correct stress assignment can be predicted without exception. In Cairene Arabic, long vowels and coda consonants contribute equivalently to both syllable weight and main stress assignment. CVV and CVC syllables, which are both treated as heavy syllables, contribute equivalently to main stress assignment in non-final position. Similarly, CVVC and CVCC syllables, which are both superheavy and appear only in final position, are treated equivalently. English patterns similarly to Cairene Arabic in that both long vowels and coda consonants contribute to weight; in other words, CVV and CVC syllables are both treated as heavy. However, long vowels are a more reliable trigger for main stress assignment than coda consonants. The presence of coda consonants in English nonce nouns has either a small effect (Domahs et al., 2004) or no effect (Guion et al., 2004) on the assignment of stress. In particular, only 20% of superheavy final syllables received main stress in English nonce word experiments (Domahs et al., 2014; Guion et al., 2003). In contrast, the presence of a long vowel was the most important predictor for stress assignment in English nonce word experiments (Guion, et al., 2003; Guion et al., 2004; Guion, 2005). If participants successfully acquire the Cairene Arabic weight system, both long vowels and coda consonants should be important predictors for correct responses, and should contribute equally to stress assignment. In contrast, if participants apply an L1 transfer strategy, long vowels should remain an important predictor for correct responses; however, the presence of coda consonants should have a small or non-significant effect on performance.

METHOD

Corpus and word selection

The words used in this experiment were drawn from the LDC Colloquial Egyptian Arabic lexicon, which consists of 51,202 words extracted from telephone conversations and dictionary entries (Kilany et al., 1997). Cairene Arabic contains a number of phonemes unfamiliar to an American English participant, such as pharyngealized consonants. The inclusion of such phonemes, particularly in words containing consonant clusters, may result in misperceptions (Davidson, 2010; Davidson & Shaw, 2012; Shaw & Davidson, 2011), making it difficult for participants to acquire the stress system in a short amount of time. Therefore, a subset of the Cairene Arabic phonemic inventory was used, with the consonant inventory restricted to segments which appear in both American English and Cairene Arabic. Vowels were restricted to the 3 short vowels /ɑ/, /ɪ/ and /ʊ/, and their corresponding long vowels. This excluded the long vowels /uː/ and /oː/ which are primarily found in words of foreign origin. Although English, unlike Cairene Arabic, does not have a phonemic vowel length contrast, it does have a distinction between tense (long) and lax (short) vowels. In both languages, short vowels pattern as light, while long vowels pattern as heavy. Tense vowels in American English contribute to stress assignment in a similar manner to long vowels in languages such as Cairene Arabic. This suggests that participants should be able to perceive the difference between long and short vowels in these stimuli. In a pilot task, a separate set of 180 American English participants were asked to transcribe the stimuli in the current experiment, and indicate which syllable was stressed. Participants correctly identified 82% of stressed syllables, suggesting that incorrect perception of stressed syllables was not a barrier to acquisition. The following sets of words were also excluded from consideration: a) one-syllable words; b) words longer than 5 syllables; c) words with multiple pronunciations; d) words of foreign origin with non-native phonology; e) words which resembled English lexical items. One syllable words were excluded because there is no logically possible variation in stress pattern. Words longer than five syllables were excluded because they were highly uncommon in the corpus, as well as due to concerns about perceptibility in longer words. The resulting subset of the original corpus consisted of 8668 words.

Stimuli

Previous experiments on stress perception or production have organized the stimuli according to one of two principles. In some experiments, words are organized according to syllable length, with shorter words being presented before longer words (Altmann, 2006; Carpenter, 2005; Carpenter, 2010; Guest et al., 2000). Other experiments organize words according to their CV-structure, describing words as sequences of consonants (C) and vowels (V); or weight, describing words as sequences of light (L), heavy (H), and superheavy (S) syllables (Domahs et al., 2014; Guions, 2005; Pater, 1997; Tremblay & Owens, 2010). For example, CVV.CVCC (HS) words may be presented separately from CV.CVCC (LS) words. This is particularly useful when examining the effect of syllable structure on performance. However, in this experiment, there were too many word types for this approach to be practical. The usable corpus contains a total of 90 different word types in terms of CV-structure (such as CVV.CVCC), or 43 different types in terms of syllable weight (such as HS). CVC and CVV syllables are both heavy (H); CVCC and CVVC syllables are both superheavy (S); this is why there is a smaller number of word types in terms of syllable weight compared to CV-structure. Rather than using CV-structure or syllable weight as an organizing principle, we organized the stimuli in this experiment into ‘constraint groups’. Words were organized into groups according to the smallest set of active OT constraints required to correctly predict stress. Each constraint group contained one or more word types, in terms of weight (such as HS or LS). No word type was associated with more than one constraint group. As a result, the number of different word types was significantly reduced.

Given the OT analysis of Cairene Arabic stress described above, stress placement may be determined by a single active constraint. For example, the constraint group 'WSP' contains words with a single, non-final heavy syllable. Knowledge of this constraint alone is sufficient to determine which syllable needs to be stressed. This constraint group contains the following word types: HL, LHL, LLHL. However, in many cases, the OT analysis cannot derive the correct output given only a single active constraint. Instead, multiple interacting constraints are necessary in order to determine correct stress placement. For example, the constraint group 'NonFin + EndRule-R' includes words with multiple heavy syllables, one of which is final. The constraint NonFin rules out stress on the final heavy syllable. The constraint EndRule-R picks the rightmost of the remaining candidates. This constraint group includes the following word types: HHH, LHHH, HHHH.

This OT-based approach can be contrasted with alternate analyses based on the surface properties of the input presented to participants. Given the words in the constraint group 'WSP', participants may observe that heavy syllables are stressed over light syllables, thus acquiring the constraint WSP and correctly deriving the output. However, participants may instead observe that these words all contain right-aligned trochees, and assume that this is the reason they receive stress. This would allow participants to correctly predict stress placement for words in the constraint group 'WSP', without acquiring the underlying OT constraint. As a result, participants would make incorrect predictions about words in constraint groups such as 'NonFin & EndRule-R & WSP', which include the constraint WSP. Inconsistent patterns of acquisition such as these may be evidence that participants are sensitive to surface stress properties, rather than OT constraint rankings. Patterns in which participants acquire a simple constraint group, such as 'WSP', but not a complex constraint group including WSP are insufficient evidence of this: an alternate explanation is simply that complex constraints are harder to acquire. However one set of inconsistent patterns does provide evidence that participants are sensitive only to surface stress patterns. These are patterns of acquisition in which participants have not acquired a simple constraint group, such as 'WSP', but perform strongly on a complex constraint group including WSP, such as 'NonFin & EndRule-R & WSP'. These patterns of acquisition should not be possible if participants' acquisition is guided by OT constraint rankings, but are consistent with an acquisition strategy based on surface stress patterns.

The words in the corpus were classified using a total of 13 constraint groups: NonFin; EndRule-R; WSP-CVV; WSP; $\text{WSP}_{\mu\mu\mu}$; EndRule-R + WSP; NonFin + EndRule-R; NonFin + WSP; Trochaic + Align-L; NonFin & EndRule-R & WSP; Align-L & EndRule-R & NonFin; Trochaic & Align-L & NonFin; Trochaic & Align-L & EndRule-R & Parse-Ft. 15 words from each constraint group were selected at random for use in the experiment. Constraint groups were ordered in terms of complexity, with groups containing a single active constraint grouped before those with 2 active constraints, and so forth. This was expected to make acquisition easier, due to the predictions of the 'less is more' hypothesis (Newport, 1990), which states that adult participants are better able to acquire a linguistic system when input is initially limited in terms of length or complexity. This had the effect of roughly sorting the words by length, as shorter words were more likely to require fewer constraints for correct stress assignment. All words used in the current experiment are presented in the Appendix.

The stimulus creation process was modelled as closely as possible on Carpenter (2005; 2010). However, Carpenter used only CV syllables, while the current experiment used 5 syllable types: CV, CVC, CVV, CVVC and CVCC. As a result, there were certain methodological differences. In Carpenter's study, a trained phonetician was used to record isolated syllables, which were then stitched together in Praat (Boersma, 2002). This method caused audible artefacts when syllables with obstruent codas were joined onto a subsequent syllable. To avoid this issue, stimuli in the current study were synthesized using the Festival synthesis software (Taylor, Black, & Caley, 1998). The default male American English voice was used, with recordings being produced at a mean of 110 Hz. The resulting stimuli were produced more consistently than would be possible with a human speaker.

The stimuli produced by Festival were segmentally accurate, but with flat and consistent affect throughout. Praat was used to manipulate the synthesized words in intensity, pitch and duration. Long vowels were created from synthesized short vowels; the adjusted duration was double that of the corresponding short vowel. For each word, every possible combination of main stress was synthesized. For example, given a 3-syllable word, 3 variants were produced: with initial, medial and final main stress respectively. Stressed syllables were systematically varied from unstressed syllables along three parameters: intensity, pitch and duration. Following Carpenter, stressed and unstressed syllables differed in intensity by 6dB; in pitch by 20%; in duration by 20%. A pitch contour was added to each word according to the position of main stress: an initial main-stressed syllable received a falling contour; a medial main-stressed syllable received a rising-falling contour; a final main-stressed syllable received a rising contour. The base duration of each syllable was varied according to its CV-structure; for example, CV syllables were assigned a shorter duration than CVCC syllables. These durations were consistent across all instances of a syllable type. For example, the duration of all unstressed CV syllables was the same; all stressed CV syllables were 20% longer. Following these manipulations, each syllable differed in terms of pitch contour in initial, medial and final position. Each syllable was identical in terms of segmental content, duration and intensity across all positions. All stimuli were vetted for naturalness by multiple native speakers of English. Glides and rhotics in coda position contained audible artefacts. As a result, these words were replaced following the vetting procedure. Some constraint groups contained very few items, making it impossible to remove these words while still selecting sufficient items for the experiment. In this case, glides and rhotics in coda position were replaced with a nasal.

Recruitment and data collection

Amazon Mechanical Turk, a crowdfunding platform which allows researchers to quickly and cheaply access a large number of participants, was used to recruit 144 participants. All participants were located in the US, and were required to have a prior approval rating of at least 95%. The experiment took place over multiple sessions on 4 consecutive days. Each session took 20-30 minutes to complete. A new session was posted every 24 hours, and expired after 24 hours. Participants were asked to wait until they had had a full night's sleep between rounds before responding. Reminders were sent to participants who had not yet completed the current round, at pre-set times throughout the day. Payment for each round was $1.50, with a $1.50 bonus for completing all 4 rounds. The retention rate was 57%, with 82 participants completing all 4 rounds. Data from a further 9 participants was excluded due to: a) technical issues (n = 5), b) being a non-native English speaker (n = 3), and c) previous exposure to a Semitic language (n = 1). Therefore, data from 73 participants was used in the final analysis.

Procedure

The experiment was created using Javascript. Data from the experiment was saved in an SQL database on a university server. Submiterator (Lassiter, 2014) was used to submit completed HITs to Mechanical Turk. Longi (Schembri, Johnson, & Demuth, 2016) was used to handle the technical issues involved in running a multi-day study: posting HITs on a regular schedule, restricting participants based on country and approval rating, ensuring that HITs on days 2-4 were available only to participants who had completed the previous day's HIT, sending reminders, approving HITs and paying bonuses.

Participants were told that they were going to learn an unknown language, and were given the opportunity to hear a sample word before the experiment began. They were asked to use headphones where practical. Data on participants' audio equipment was collected at the end of the experiment. The procedure was closely modelled on Carpenter (2005; 2010). The experiment consisted of three phases: familiarization, training and generalization. Figure 1 presents a visual representation of the familiarization phase; Figure 2 represents the training and generalization phases.

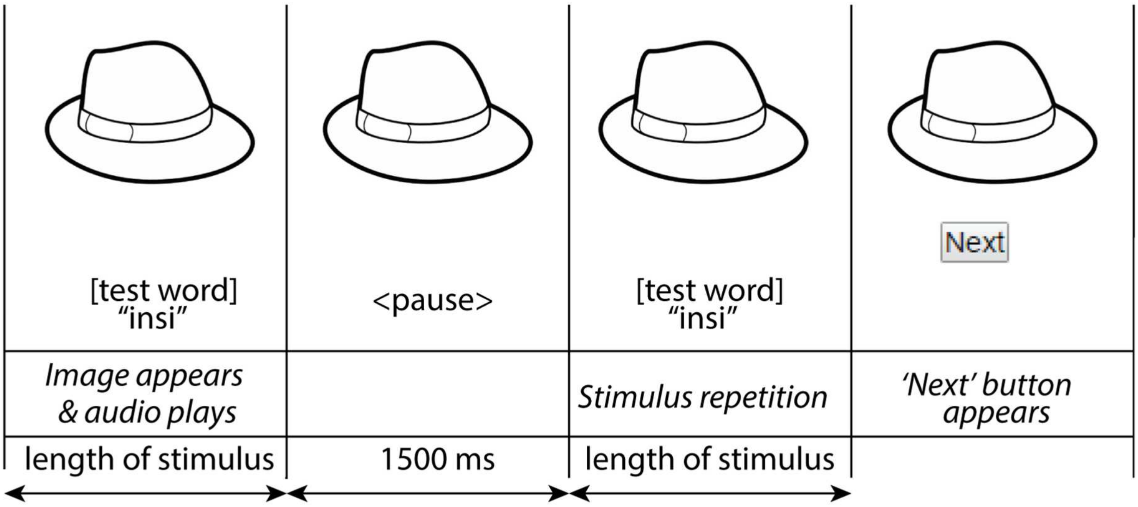


Figure 1: Familiarization phase

During familiarization, participants listened to each word twice. Each word was paired with an image, following Carpenter's findings that the use of images improved word learning. This had the additional effect of contextualizing the words as nouns.

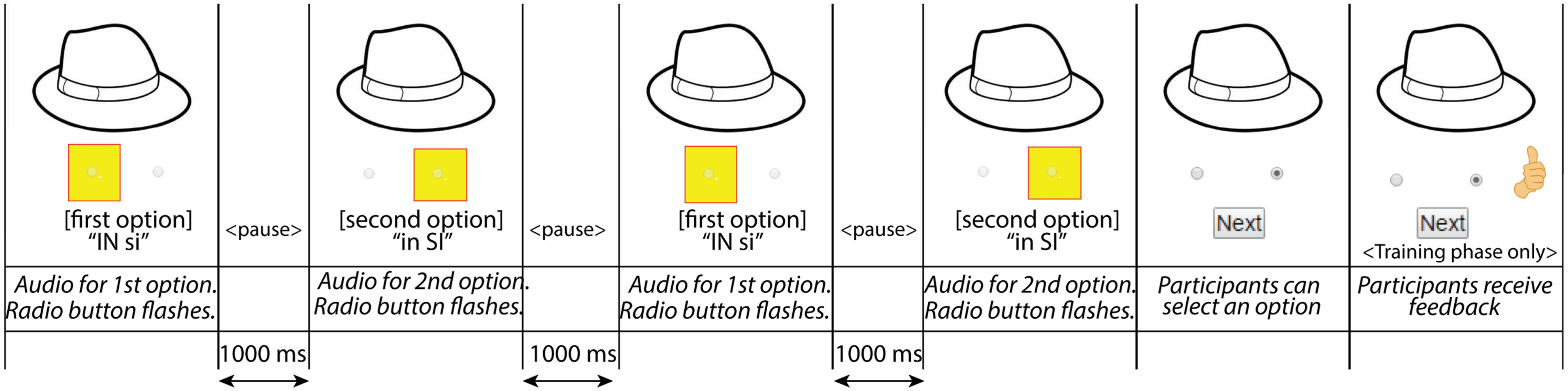


Figure 2: Training and generalization phase. Note that the image remains on screen throughout. It has been removed here for reasons of space. The use of capital letters in this diagram corresponds to an auditorily stressed syllable.

During training, participants were tested on the words they had heard in familiarization through a two-alternative forced choice task. The two choices were ordered randomly, and consisted of the correctly stressed word, as well as a randomly selected incorrectly stressed word. For example, given a 4 syllable word, the incorrect choice presented to participants was randomly selected from the 3 possible alternatives. Participants were instructed to select the word which sounded most similar to the stimulus presented during familiarization. Each alternative was presented twice. Once the sound files stopped playing, participants were given 5 seconds to answer. If they did not answer within the time limit, the experiment moved on to the next question. Participants were given feedback on their response; the software displayed a thumbs-up image for a correct answer, and a thumbs-down image for an incorrect answer. Generalization was identical to training, with two exceptions: a) participants were tested on words they had not heard previously, and b) participants were not given feedback on their response. Before the generalization phase began, participants were informed that they would encounter words which were unfamiliar to them. They were asked to select whichever option sounded most like a word in the language they had been learning.

The first three days of the experiment followed a similar structure; this is represented in Table 1. Participants were presented with alternating familiarization and training blocks. During training, participants were tested on the words encountered in the immediately preceding familiarization block. A review block was presented for every two familiarization and training blocks. Each review block contained both familiarization and training for each word encountered so far. After all familiarization, training and review blocks were completed, participants were asked to complete a generalization block. This block contained words which had not been seen before; each word in the familiarization block was matched with a novel word which was identical in terms of

CV-structure. The set of stimuli presented within each block was the same for each participant; however, presentation order was randomized within the block.

Block	Description
Familiarization Block 1	Subjects hear 5 training words from a constraint group.
Training Block 1	Subjects are tested on the 5 words in Familiarization Block 1.
Familiarization Block 2	Subjects hear 5 training words from a new constraint group.
Training Block 2	Subjects are tested on the words in Familiarization Block 2.
Review Block 1	Subjects complete a familiarization and training block for all items presented so far (10 items).
Familiarization Block 3	Subjects hear 5 training words from a new constraint group.
Training Block 3	Subjects are tested on the words in Familiarization Block 3.
Familiarization Block 4	Subjects hear 5 training words from a new constraint group.
Training Block 4	Subjects are tested on the words in Familiarization Block 4.
(Familiarization Block 5)	***This block was present on Day 1 only.*** This is because there are 13 constraint groups (an odd number). On Day 2 and Day 3, the experiment moved on to Review Block 2.
(Training Block 5)	***This block was present on Day 1 only.*** This is because there are 13 constraint groups (an odd number). On Day 2 and Day 3, the experiment moved on to Review Block 2.
Review Block 2	Subjects complete a familiarization and training blocks for all items presented so far. 25 items are reviewed on Day 1. 20 items are reviewed on Day 2 and Day 3.
Generalization Block	Subjects are tested on previously unseen words. Each word in the current day's Familiarization and Training blocks is matched with a novel word which is identical in terms of CV-structure. 25 items are presented on Day 1. 20 items are presented on Day 2 and 3.

Table 1: Experiment structure for days 1-3. Note that days 2 and 3 began with an additional Review Block (not shown), in which 1 word was included from each constraint group presented from the previous day(s).

Each day, participants were exposed to words in different constraint groups. Table 2 illustrates the different stimuli that were presented on each day. The familiarization, training and review phase included 5 words from each constraint group; 5 novel words from each constraint group were presented in generalization. Of the 13 constraint groups in total, participants learned words from 5 different constraint groups on day 1; from 4 groups on day 2; from 4 groups on day 3. Simple constraint groups were ordered before complex constraint groups. For example, words in the constraint group NONFIN were taught on day 1; words in the constraint group TROCHAIC & ALIGN-L & ENDRULE-R & PARSE-FT were taught on day 3. Days 2 and 3 began with a review of the words learned on the previous days. Day 4 was devoted exclusively to generalization and consisted of a single, large generalization block. Participants were presented with 5 unfamiliar words from all 13 constraint groups. Participants' performance on this day is therefore an important measure of overall acquisition: in order to be successful, participants must remember the patterns acquired throughout the entire experiment. Participants completed each experiment over the course of 20-30 minutes.

Day 1	Day 2	Day 3	Day 4
1. WSP$_{\mu\mu\mu}$ 2. WSP 3. NonFin 4. EndRule-R 5. WSP-CVV	1. EndRule-R & WSP 2. NonFin & EndRule-R 3. NonFin & WSP 4. Trochaic & Align-L	1. NonFin & EndRule-R & WSP 2. Align-L & EndRule-R & NonFin 3. Trochaic & Align-L & NonFin 4. Trochaic & Align-L & EndRule & Parse-Ft	5 novel words from all 13 constraint groups

Table 2: Constraint groups presented on each day. On days 1-3, 5 words from each constraint group were presented in the familiarization and testing phase; a further 5 words were presented in the generalization phase.

RESULTS

Figure 3 presents correct responses for both training and generalization on each day of the experiment, averaged across participants. Note that participants received no training on day 4.

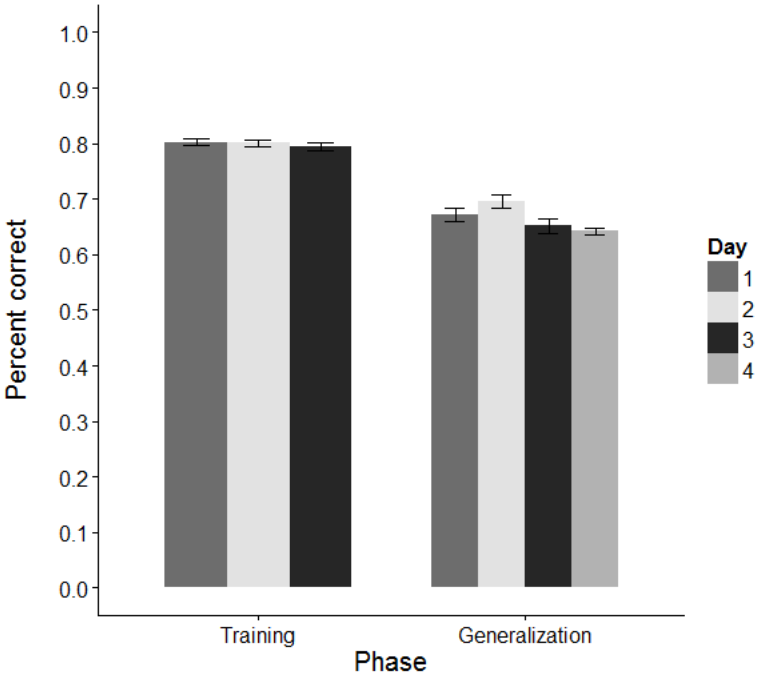


Figure 3: Correct responses averaged over participants. Error bars represent the standard error.

Participants scored an average of 80% correct in the training phase across all days. They scored between 64-70% in the generalization phase. This is comparable to Carpenter's (2005, 2010) studies, in which participants averaged 89-90% correct in training, and 61-70% in generalization, depending on condition. Overall, participants were able to acquire a complex stress system over the course of multiple days, in a manner comparable to the acquisition of a simpler stress pattern over a single day.

Successful acquisition is measured through participants' ability to generalize to novel stimuli. A generalized linear mixed-effects model was fitted using the lme4 package in R (Bates et al., 2015), which was also used to compute p-values. The model included individual participant responses as the dependent variable, coded as correct (1) or incorrect (0). The following fixed effects were included in the model: Constraint Name, and Experiment Day (1-4). No significant interactions were found. Subjects and items were entered as random variables with random intercepts. The formula for the model was Correct ~ Constraint_Name + Day + (1 | Worker_ID) + (1 | Item_Number). Factors were dummy coded, with WSP-CVV as the reference category for the factor Constraint_Name, and day 1 as the reference category for the factor Day. Table 3 presents all main effects with coefficient estimates, standard error, z- values and p-values. A supplementary model to check for order effects is included in Appendix A.

	Coefficient	*SE*	*z*	*p*	
(Intercept)	0.775	0.154	5.050	< 0.001	***
EndRule	0.358	0.189	1.893	0.058	.
EndRule & WSP	0.625	0.198	3.165	0.002	**
NonFin	-0.434	0.186	-2.331	0.020	*
NonFin & EndRule	0.548	0.185	2.965	0.003	**
NonFin & EndRule & WSP	0.609	0.191	3.184	0.001	**
NonFin & WSP	0.371	0.187	1.981	0.048	*
Trochaic & A-L	-0.163	0.184	-0.888	0.374	
Trochaic & A-L & EndRule	0.054	0.189	0.286	0.775	
Trochaic & A-L & Endrule & Nonfin	0.200	0.191	1.045	0.296	
Trochaic & A-L & NonFin	-0.319	0.178	-1.787	0.074	.
WSP	0.028	0.182	0.153	0.879	
WSPµµµ	0.198	0.186	1.064	0.287	
day2	-0.249	0.125	-1.997	0.046	*
day3	-0.235	0.123	-1.918	0.055	.
day4	-0.268	0.085	-3.168	0.002	**

Table 3: Constraints and Experiment Day with coefficient, standard error, z values and p values

The model demonstrates that participants performed best on the first day, with significantly worse performance on days 2 (z = -1.997, p = 0.046) and 4 (z = -3.168, p = 0.002). A negative performance effect was also found for day 3, but this was only approaching significance (*z* = -1.918, *p* = 0.055). This is consistent with the organization of stimuli in order of complexity: participants were better able to acquire the simple stimuli presented on day 1 compared to the more complex stimuli presented on subsequent days. Participants performed significantly better than average on words in the following constraint groups: EndRule & WSP (*z* = 3.165, *p* = 0.002), NonFin & EndRule (*z* = 2.965, *p* = 0.003), NonFin & EndRule & WSP (*z* = 3.184, *p* = 0.001), and NonFin & WSP (*z* = 1.981, *p* = 0.048). Participants performed worse than average on words in the constraint group NonFin (*z* = -2.331, *p* = 0.2). No significant effects were found for words in the remaining constraint groups.

One interpretation of these results is that participants were able to learn certain constraint rankings better than others: this would explain the disparity in performance between words in different constraint groups. However, a close examination of the constraints involved reveals an inherent contradiction with this interpretation. For example, knowledge of the constraint group NonFin & EndRule & WSP implies that participants have acquired an understanding of the three individual constraints, as well as the ranking between them. However, participants have not acquired words in the constraint group NonFin: participants' performance on these words is significantly worse than on any other constraint group. The constraint NonFin penalizes any word with final main stress. The constraint group NonFin contains words in which NonFin is the only active constraint; that is, all other constraints which are relevant to the assignment of main word stress are dominated by NonFin. There are 2 word types in the constraint group NonFin: LH (such as 'za.mat) and HH (such as 'ma:.ziz). These words provide the simplest possible illustration of the principle of nonfinality: participants can either choose to stress the initial syllable, obeying NonFin, or to stress the final syllable, violating NonFin. Participants in the current experiment performed poorly on these words; that is, they chose to stress the final syllable. This indicates that participants may not have learned that the constraint NONFIN is highly ranked. Similarly, the model shows that the remaining individual constraints were not fully acquired. No significant effect was found for the constraint group WSP, suggesting that participants do not consistently make use of this constraint. Participants do perform better than average on words in the constraint group ENDRULE, but this effect is only approaching significance ($z = 1.893$, $p = 0.058$). This problem affects all constraint groups for which the model finds a positive significant result: all of these groups contain some combination of the individual constraints NONFIN, ENDRULE and WSP.

Based on the observations above, it appears that participants have not truly acquired the constraint groups indicated by the model. However, participants do perform better on these words. A closer look at the words contained in these constraint groups reveals that they all contain a stressed heavy penultimate syllable. This provides an alternate explanation for participants' performance. In Cairene Arabic, heavy penultimate syllables are always stressed, while in English they are stressed almost categorically. This can be analysed as an L1 transfer effect: both the native and target language predict the correct stress assignment; hence, these words are easier to acquire. This observation suggests the possibility of an alternate analysis based on a comparison between the predictions of English and Cairene Arabic for stress assignment. Both languages make similar predictions in the case of a heavy penultimate syllable. However, when a word does not contain a heavy penultimate syllable, the English and Cairene Arabic word stress system differ in their distribution of stress. Initial stress is the default pattern for English nouns, but is far less common in Cairene Arabic. Final stress is highly uncommon in English nouns, but occurs more frequently in Cairene Arabic. In addition, the two stress systems differ in their treatment of weight. In Cairene Arabic, the weight system is deterministic: correct stress assignment is entirely predictable from the sequence of light, heavy and superheavy syllables in a word. The English stress system has some degree of weight sensitivity; however, this is probabilistic in nature. In Cairene Arabic, CVV and CVC syllables are designated as heavy, and contribute equivalently to main stress assignment in non-final position. In English, while CVV and CVC syllables are both heavy, CVV syllables trigger main stress assignment more often than do CVC syllables (Guion, et al., 2003; Guion et al., 2004; Guion, 2005).

A second analysis was performed, using these differences between the two stress systems to shed further light on participants' patterns of acquisition. Cairene Arabic and English differ in their distribution of initial and final stress; therefore, the factor 'Stress Position' was added to the model, distinguishing between words with initial, medial or final stress. Words with final stress are exclusively contained within 2 constraint groups. Words ending in a superheavy syllable are contained in the constraint group WSPμμμ, which implies the ranking WSPμμμ >> NonFin. Words ending in a CVV syllable are contained in the constraint group WSP-CVV, which implies the ranking WSP-CVV >> NonFin. The simple relationship between final stress position and these constraint groups made it feasible to separate out the effects of these constraint rankings on participants' performance. Following this step, the factor 'Stress Position' contained four levels: initial, medial, final (WSPμμμ) and final (WSP-CVV). An additional factor was added to account for the effect of weight on participants' performance, containing two levels: light and heavy. Note that superheavy syllables are all contained within the constraint group WSPμμμ, and are thus already accounted for. Although Cairene Arabic treats CVV and CVC syllables equivalently in non-final position, participants may have a tendency to stress CVV syllables over CVC syllables due to L1 transfer. In order to account for this possibility, an additional factor 'Long Vowel' was added, denoting the absence (coded as -1) or presence (coded as 1) of a long vowel in a word. This enables the model to distinguish whether participants assign stress: 1) equally to all syllables regardless of weight, 2) more often to all heavy syllables than light syllables, or 3) more often to CVV syllables than light or CVC syllables. The formula for the model was Correct ~ Stress_Position + Weight + Long_Vowel + Day + (1 | Worker_ID) + (1 | Item_Number). The reference category for the factor Stress_Position was Medial; for the factor Weight was Light; for the factor Long_Vowel was long_vowel0 (absence of long vowel).

Model comparison between the OT constraint model described in Table 3, and the current model based on surface stress properties, was carried out to determine optimal model fit. Table 4 presents results from a likelihood ratio test comparing the two models.

	Df	*AIC*	*BIC*	*LogLik*	*Deviance*	*Chisq*	*Chi Df*	*Pr(>Chisq)*
stress_pos_model	11	13805	13886	-6891.6	13783			
constraint_model	18	13811	13943	-6887.6	13775	7.9712	7	0.3351

Table 4: Likelihood ratio test comparing the OT-constraint model with the surface stress model.

Table 4 demonstrates that the surface stress model is simpler than the OT constraint model, with 7 fewer degrees of freedom. These results suggest that the OT constraint model does not provide a significantly better fit than the surface stress model (*p* = 0.3351). However, the likelihood ratio test is designed for use with nested models. The measures AIC and BIC are designed to compare non-nested models, and may therefore provide a more appropriate measure of model fit in this instance. A lower value for both AIC and BIC indicate a better model fit. Both AIC and BIC demonstrate that the surface stress model fits the data better than the OT constraint model. However, the two models score very similarly on all metrics, suggesting that the surface stress model's relative simplicity is the deciding factor. This conclusion is supported by the observation that the surface stress model scores particularly well on BIC, which penalizes model complexity to a greater extent than AIC. Table 5 presents all main effects for the surface stress model with coefficient estimates, standard error, z values and p-values. A supplementary model to check for order effects is included in Appendix A.

	Coefficient	*SE*	*z*	*p*	
(Intercept)	0.966	0.142	6.789	< 0.001	***
stress_posINITIAL	-0.563	0.102	-5.536	< 0.001	***
stress_posFINAL_CVV	-0.498	0.157	-3.169	0.002	**
stress_posFINAL_μμμ	-0.235	0.201	-1.170	0.242	
long_vowel1	0.218	0.108	2.028	0.043	*
weightheavy	0.062	0.112	0.558	0.577	
day2	-0.117	0.115	-1.020	0.308	
day3	-0.201	0.118	-1.706	0.088	.
day4	-0.218	0.082	-2.663	0.008	**

Table 5: Stress Position, Weight, Long Vowel and Day: coefficient, standard error, z and p values

The model revealed that, in agreement with the first model, participants performed best on words with medial main stress. However, this model provides additional insights: in comparison, participants' performance was worse on words with initial main stress ($z = -5.536$ $p < 0.001$). Participants' performance on words with final stress differed depending on constraint group. Participants performed poorly on words ending with a CVV syllable ($z = -0.498$ $p = 0.002$);

however, no significant effect was found for words ending with a superheavy syllable. Participants were more likely to stress words containing a long vowel (*z* = 2.028, *p* = 0.043); once this effect was taken into account, heavy syllables had no significant effect on participants' performance. This means that participants were more likely to stress CVV syllables over CVC and CV syllables. When adjusted for the above factors, participants' performance was best on the first day of the experiment. This result was similar to the results seen in the first model in Table 3; however, while both models found that overall performance was best on day 1 and worst on day 4, the two models differed in their conclusions for days 2 and 3. In the current model, significant effects were found for worse performance on day 3 (*z* = -1.974, *p* = 0.048) and day 4 (*z* = -2.831, *p* = 0.005); no significant effects were found for day 2.

Main stress position was a significant factor affecting participants' performance. As demonstrated in the statistical model in Table 5, participants performed best on words with medial stress (70% correct), followed by final stress (66%) and initial stress (56%). The category 'medial stress' includes words with non-final stress on both the second and third syllable, as there was no difference in performance between these two groups. Participants performed significantly worse on words with final stress in the constraint group WSP-CVV (64%) as compared to those in the constraint group WSP μμμ (68%). Participants' performance by main stress position is illustrated in Figure 4 below.

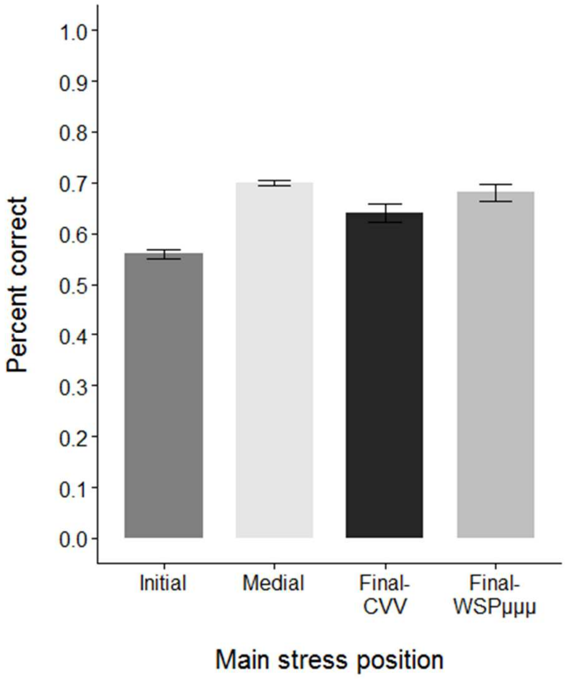


Figure 4: Correct responses by main stress position. Error bars represent standard error.

GENERAL DISCUSSION

Overall, participants were successful in acquiring the complex Cairene Arabic stress system. However, the statistical models in Tables 3 and 5 demonstrate that there are systematic patterns in the types of stimuli that were more or less readily acquired. Participants' performance was significantly affected by syllable weight, main stress position and experiment day.

Syllable weight was a significant factor affecting performance; however, the weight system acquired by participants differed from that of the underlying Cairene Arabic system. In Cairene Arabic, heavy syllables (CVC, CVV) are preferentially stressed over light syllables (CV); superheavy syllables (CVCC, CVVC) are preferentially stressed over heavy syllables. Participants were significantly more likely to stress words containing a long vowel; that is, they were more likely to stress CVV and CVVC syllables than CVC and CVCC syllables respectively. However, after taking this effect into account, there were no significant effects for stressing light, heavy and superheavy syllables. In other words, participants were equally likely to stress light, heavy and superheavy syllables which did not contain a long vowel; that is CV, CVC and CVCC. This result means that a fundamental aspect of the Cairene Arabic stress system was not acquired. These observations can be explained as an L1 transfer effect. In English, a syllable containing a long vowel is significantly more likely (60%) to receive main stress than a syllable containing a short vowel (35%), according to data from the CELEX database (Guion, 2003). The absence or presence of coda consonants has a smaller effect on stress assignment. Corpus data for the effect of coda consonants on stress assignment is not available. However, participants in novel word experiments are significantly more likely to stress a CVV syllable over a CVC syllable, and a CVVC syllable over a CVCC syllable (Domahs, 2014; Guion et al., 2003; Guion et al., 2004; Guion, 2005). These patterns mirror the results seen in the current experiment, which can therefore be analyzed as a result of L1 transfer. This tendency may be due to the importance of duration as a cue for stress in English. Unstressed vowels in English, unlike in Arabic, are reduced in duration as well as quality. In comparison, long vowels may be especially salient for English speakers.

Participants' performance is worst overall on words which should receive initial stress. These results demonstrate that participants are not applying an L1 strategy to words with initial stress. Both the L1 and target language predict initial stress, yet participants perform worst overall on these words. There remains a possible confounding factor: 92% of English bisyllabic nouns receive initial stress, but longer words are less likely to be initially stressed (Clopper, 2002). It is possible that participants correctly assign initial stress more consistently in bisyllabic words, reflecting the L1 pattern. However, word length, as well as the interaction of word length and syllable structure, were not found to be significant factors. A supplementary model to demonstrate this is included in Appendix A. The results for bisyllabic words mirror the overall results. This is consistent with participants' poor performance on words in the constraint group NONFIN, which contains only initially stressed 2-syllable words. English 2-syllable nouns are overwhelmingly stressed initially, and stress assignment in English generally obeys the principle of nonfinality: L1 transfer effects should therefore favor initial stress assignment. This is a case in which learners produce interlanguage forms which are unlike both the native and target language, similar to patterns seen in previous research on the L2 acquisition of word stress. One possible explanation is that this is an anti-transfer effect. As the task asks participants which of two options is most likely to belong to an unfamiliar language, a reasonable strategy may be to pick the least English-like option, and thereby avoid initial stress. However, participants' overall strategy more closely resembles an L1 transfer strategy, in which participants are more likely to select words which conform to English patterns. For example, participants are most successful on words with heavy penultimate syllables and words containing long vowels, which receive stress in both English and Cairene Arabic. If participants were applying a broad anti-transfer strategy, the opposite pattern would be expected. It is unclear why participants would apply such a strategy only for words with initial stress, but not elsewhere. Another explanation for the observed behaviour is overgeneralization. Cairene Arabic words are stressed initially at a much lower rate than English nouns. In the current experiment, 43% of 2 and 3 syllable stimuli receive initial stress, while longer words cannot be initially stressed in Cairene Arabic. Participants observe the relative lack of initial stress in the input, and overapply this knowledge, thus avoiding initial stress even where appropriate. Similar effects are reported in the literature. For example, late Spanish-English bilinguals produce and prefer initial stress in English in contexts where this is incorrect in both Spanish and English. Guion et al. (2004) argue that this is overgeneralization of the English tendency towards initial stress. Similarly, Lord (2001) notes that the majority of errors made by English learners of Spanish involve overgeneralization of the default penultimate stress location.

Participants' performance on words with final stress exhibits a number of interesting patterns. Participants correctly assign stress to words in the constraint group WSPμμμ more often than to those in the constraint group WSP-CVV. This is evidence of differential rates of acquisition for these constraints. This is particularly interesting because in novel word experiments, English speakers tend not to stress a final superheavy syllable (Domahs et al., 2014; Guion et al., 2003). Because of this, our prediction was that participants may perform poorly on words in the constraint group WSPμμμ due to L1 transfer effects. This prediction is not borne out: no significant difference was found between participants' performance on words with medial stress, which was best overall, and words in the constraint group WSPμμμ. Participants' poor performance on words in the constraint group WSP-CVV is puzzling. Words ending in a CVV syllable tend to attract stress in English; see, for example 'allow', 'review'. However, to our knowledge, there is no data on participants' performance on these words in novel word experiments; this suggests a possible avenue for future research.

When comparing performance between the training and generalization phase, words with final stress exhibit some unusual patterns. In the training phase, participants performed significantly better on words with final stress (86% correct), as compared to medial stress (81% correct). Participants' divergent performance on words with final stress in the separate constraint groups is even more pronounced in training: participants perform significantly better on words in the constraint group WSPμμμ (89%), as compared to WSP-CVV (82%). This is illustrated in Figure 5.

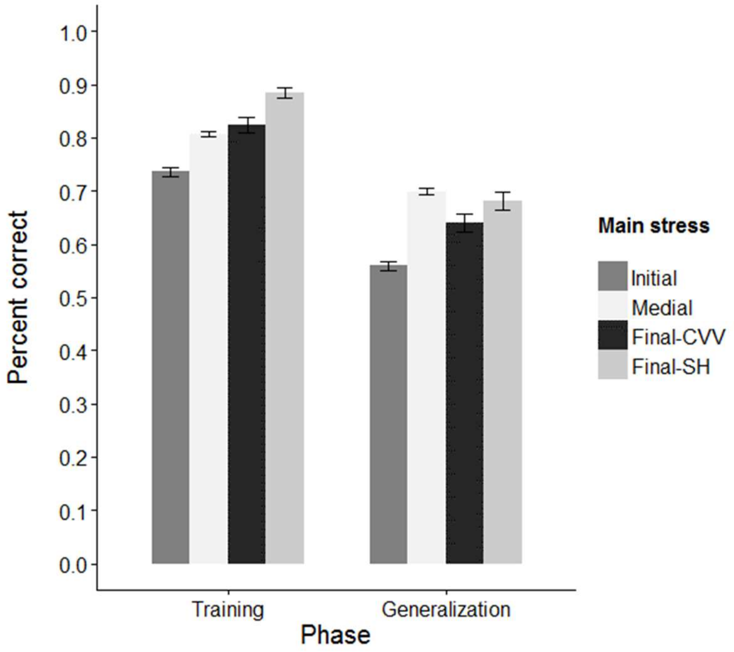


Figure 5: Correct responses by phase (training or generalization) and main stress position. Error bars represent standard error.

Participants in ALL experiments are known to reflect the patterns observed in the input more accurately in training than generalization. This is due to the memory effect. The training phase occurs directly after the familiarization phase. This means that participants have heard items being modelled correctly only minutes earlier. As a result, it is possible for participants to remember the correct answer. During the generalization phase, novel items are presented. These items are unfamiliar, so participants cannot use memory to answer correctly. Therefore, it is much more difficult for participants to answer correctly in the generalization phase than in the training phase. This is despite the fact that the same task is used for both the training and generalization phase. As expected, participants consistently perform better in training than generalization: this effect is seen throughout. However, overall trends should remain the same in training and generalization. For example, participants perform better on words with long vowels than words with no long vowels, in both the training and generalization phase. However, Figure 5 illustrates a case in which the direction of an effect is reversed in generalization: in the training phase, participants perform better on words with final stress compared to medial stress; in the generalization phase, participants perform better on words with medial stress compared to final stress. These observations hold true for all words with final stress: those in the constraint group WSPμμμ, as well as those in the constraint group WSP-CVV. A model was fitted to the data to investigate whether this training effect on main stress position is significant, adding an interaction between phase (training or generalization) and each of the linguistic factors in the model. The formula for the model was Correct ~ Phase * Long_Vowel + Phase * Weight + Phase * Stress_Pos + Day + (1 | Worker_ID) + (1 | Item_Number). The reference category for the factor Phase was Training. Table 5 presents all main effects with coefficient estimates, standard error, z-values and p-values.

	Coefficient	*SE*	*z*	*p*	
(Intercept)	1.497	0.150	9.956	< 2e-16	***
phase_generalization	-0.662	0.126	-5.250	< 0.001	***
long_vowel1	0.209	0.111	1.889	0.059	.
stress_posINITIAL	-0.411	0.113	-3.633	0.000	***
stress_posFINAL_CVV	-0.015	0.188	-0.082	0.935	
stress_posFINAL_μμμ	0.568	0.216	2.628	0.009	**
weightheavy	-0.011	0.126	-0.086	0.932	
day2	0.029	0.081	0.354	0.724	
day3	-0.022	0.092	-0.236	0.814	
day4	-0.119	0.069	-1.718	0.086	.
phase_generalization: long_vowel1	0.011	0.147	0.076	0.939	
phase_generalization: stress_posINITIAL	-0.122	0.140	-0.871	0.384	
phase_generalization: stress_posFINAL_CVV	-0.443	0.230	-1.924	0.054	.
phase_generalization: stress_posFINAL_μμμ	-0.720	0.265	-2.720	0.007	**
phase_generalization: weightheavy	0.112	0.150	0.743	0.457	

Table 6: Training and generalization: coefficient, standard error, z values and p values

As expected, participants performed best in the training phase, with accuracy dropping in the generalization phase (z = -0.662, p < 0.001). There was no significant interaction of phase (training or generalization) with words containing a stressed heavy syllable or long vowel, or words with initial stress. However, there was a significant interaction between phase and words containing final stress: participants were significantly more likely to perform well on words in the constraint group WSPμμμ (z = -0.720, p = 0.007) in generalization, as compared to training. A similar effect was found for words in the constraint group WSP-CVV, although this effect was only approaching significance (z = -0.443, p = 0.054). In training, participants' performance on words in the constraint group WSPμμμ was close to ceiling (89% correct), yet performance dropped off sharply in the generalization phase. The statistical model demonstrates that this is unusual, and that other linguistic factors do not experience this effect. Participants were able to correctly perceive and categorize words in the constraint group WSPμμμ during training. One hypothesis is that these words were later recategorized due to the effect of the L1, in which word-final superheavy syllables are rarely stressed. However, further research is required to determine whether this is the case.

The methodology described in this paper is particularly well suited to investigating the learning of languages which are understudied due to a lack of available participants. Aspects of the methodology used in Artificial Language Learning (ALL) and L2 acquisition research were combined in order to study beginning L2 acquisition in a controlled manner. Participants begin the experiment with no prior knowledge of the target language, making recruitment much easier: researchers do not have to locate multiple participants who are already learning an understudied language. Instead, participants who have experience with similar languages must be excluded, a vastly simpler task. The use of Mechanical Turk as a platform for these experiments further facilitates recruitment. Mechanical Turk allows researchers to quickly access a large, diverse population of potential participants at a low cost, enabling large-scale studies. Participants can be automatically selected on the basis of their country of residence, and researchers can manually filter participants based on language background, ensuring that the participant pool is sufficiently homogenous. Facilitating research on a wider range of languages is especially relevant to the study of main word stress, as previous research in this area has almost exclusively focussed on the acquisition of English or other European languages. However, previous ALL research demonstrates that many other aspects of language acquisition, such as the acquisition of morphological or syntactical structures, can be studied in this way. The results in the current experiment demonstrate that participants' acquisition is comparable to that achieved in other ALL experiments. Additionally, we present evidence of patterns of acquisition described in the L2 acquisition literature, such as L1 transfer and overgeneralization, indicating that this is a viable methodology for studying L2 acquisition.

CONCLUSION

Optimality Theory constraints were used to organize the stimuli in the current experiment into a manageable number of word types. These groupings were designed to demonstrate variable acquisition rates for each constraint group, however, this proved problematic. Positive significant effects were found for 5 constraint groups; no significant effects were found for a further 6 constraint groups; a negative significant effect was found for the constraint group NonFin, demonstrating that participants' performance on words in this constraint groups was worst overall, compared to all other constraint groups. Acquisition of certain constraint groups implies that others have been previously acquired. For example, acquisition of the constraint group NonFin & EndRule implies that knowledge of both of the individual constraints (NonFin and EndRule), as well as the ranking between them, has been acquired. Positive significant effects were found for 5 constraint groups; however, these all implied prior acquisition of constraint rankings for which no significant effect, or a significant negative effect was found. As a result, it was not possible to conclude that participants had truly acquired these constraints and constraint rankings.

The constraint groups which displayed a positive significant result all contained only heavy penultimate syllables; these are always stressed in Cairene Arabic, and almost categorically in English. This observation allowed for an alternate analysis based on the differences and similarities between the English and Cairene Arabic stress system: in terms of main stress placement, and the underlying weight system. On the surface, the distribution of stresses in the word is the clearest difference between the English and Cairene Arabic word stress system: English nouns tend to favour initial stress, and avoid final stress; Cairene Arabic words occur with initial stress less often, and with final stress more often. Beyond these surface properties, the English and Cairene Arabic stress systems differ on a more fundamental level. The Cairene Arabic system is driven by its weight system, such that stress placement can be predicted when the sequence of light, heavy and superheavy syllables in a word is known. The English stress system exhibits some degree of weight sensitivity; however, this is a probabilistic pattern that has only a small effect on English stress. Participants' knowledge of the underlying weight system demonstrates the effect of L1 transfer. Heavy syllables in Cairene Arabic are treated identically except in word-final position; however, in English, CVV syllables are more likely to receive stress than CVC syllables. Participants' performance in the current experiment mirrors the English, rather than Cairene Arabic, stress system: participants are significantly more likely to stress words containing a long vowel (CVV), but there is no significant effect for words containing a heavy syllable with no long vowel (CVC). Participants performed significantly better on words in the constraint group WSPμμμ, which demonstrated the ranking WSPμμμ >> NONFIN than on words in the constraint group WSP-CVV, which demonstrated the ranking WSP-CVV >> NONFIN. However, the reason for this differential acquisition is unclear. Research on American English participants' stress assignment strategies for CVV-final words is unavailable. This suggests an avenue for future research, which may shed light on whether this behaviour is affected by the L1. Finally, evidence of overgeneralization is seen in participants' over-avoidance of initial stress, leading to distributional patterns seen in neither the L1 nor the target language.

APPENDIX A: SUPPLEMENTARY STATISTICS

Descriptive Statistics

Each of the factors included in the statistical models in Table 3, 5 and 6 are illustrated through bar graphs, plotted against accuracy, below.

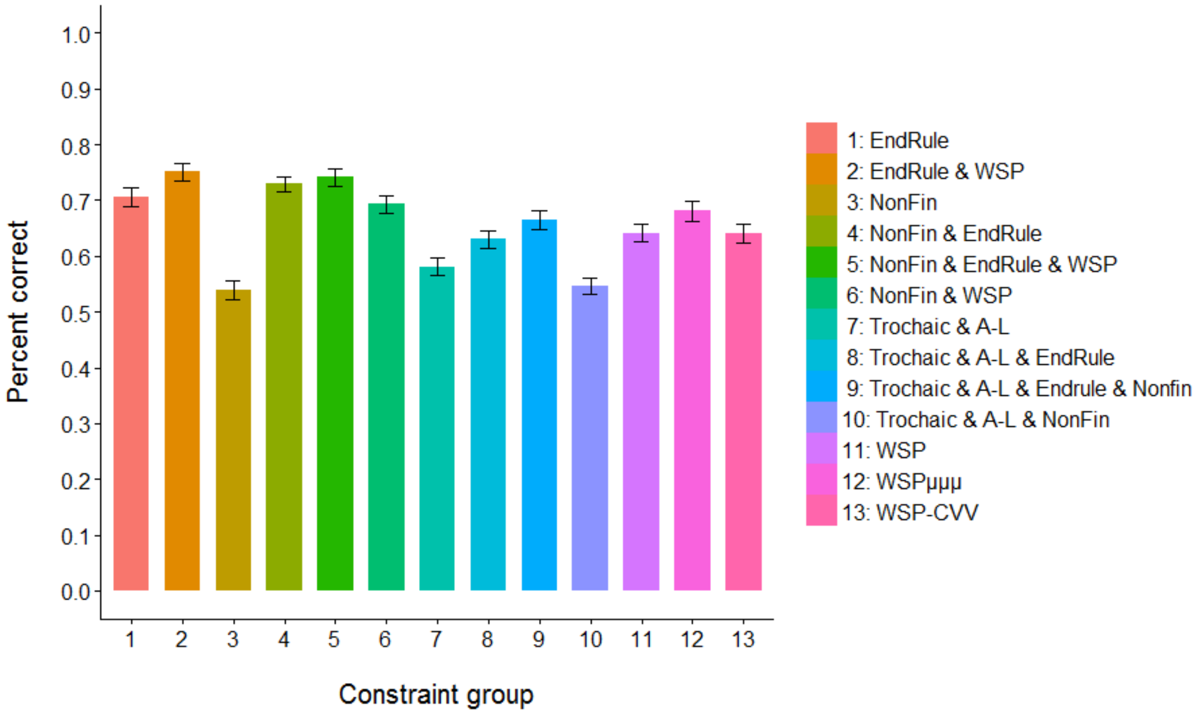


Figure A1: Correct responses by constraint group. Error bars represent standard error.

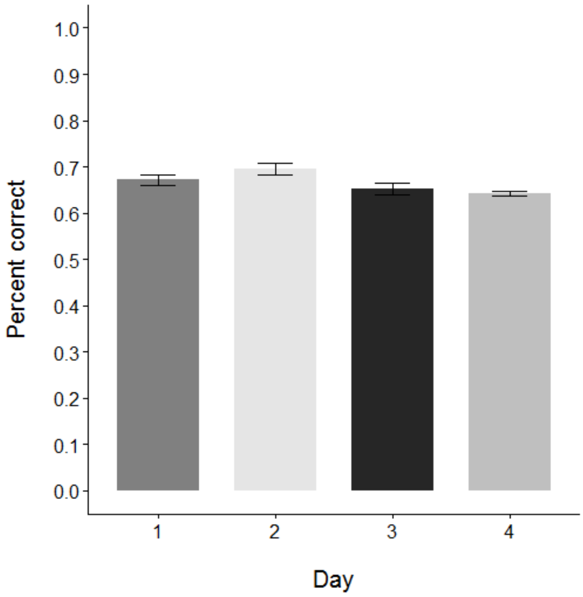


Figure A2: Correct responses by day. Error bars represent standard error.

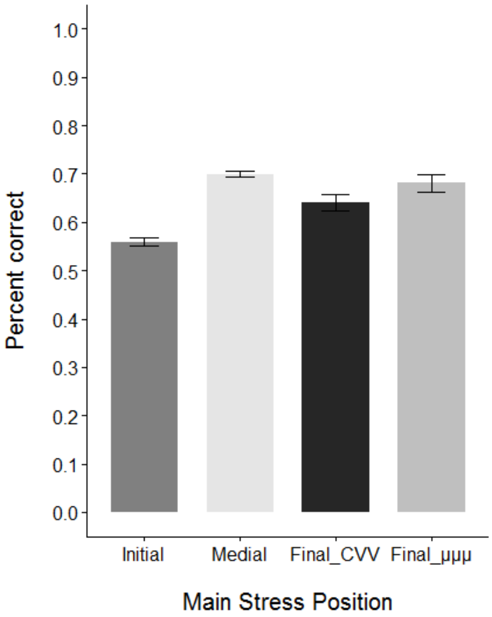


Figure A3: Correct responses by main stress position. Error bars represent standard error.

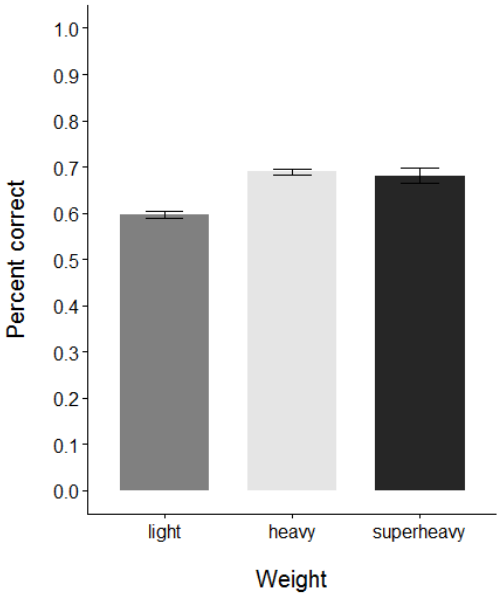


Figure A4: Correct responses by weight. Error bars represent standard error.

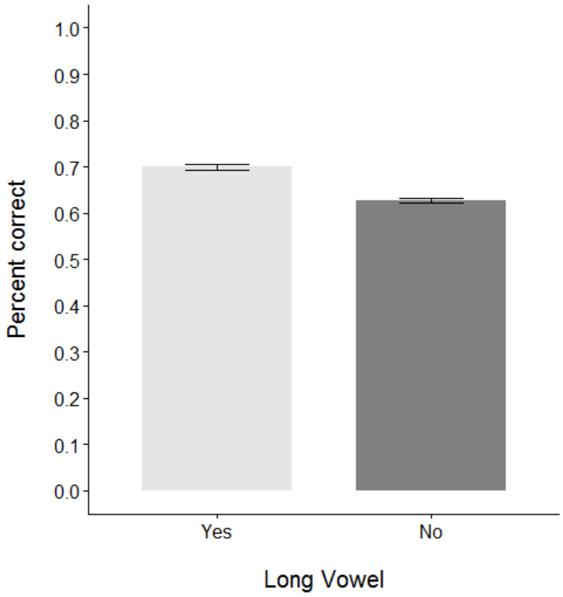


Figure A5: Correct responses by presence of long vowel. Error bars represent standard error.

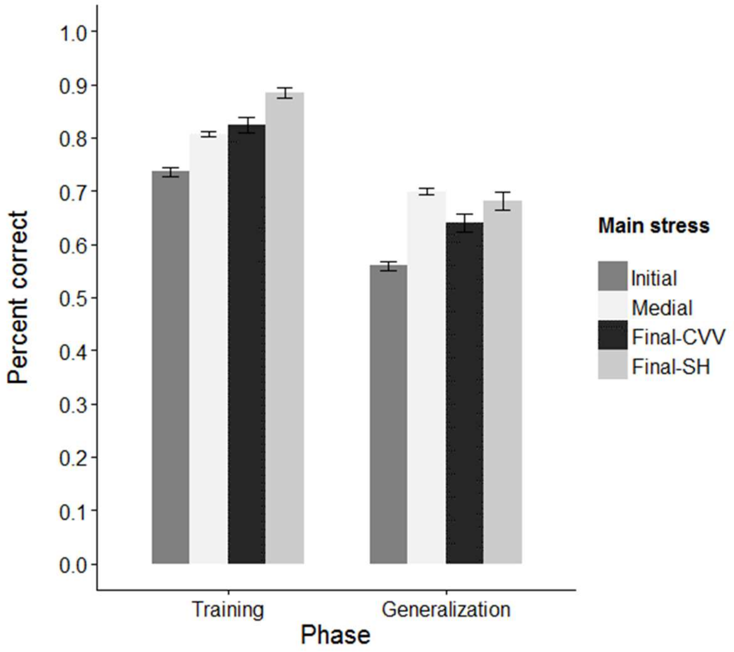


Figure A6: Correct responses by phase (training or generalization) and main stress position. Error bars represent standard error.

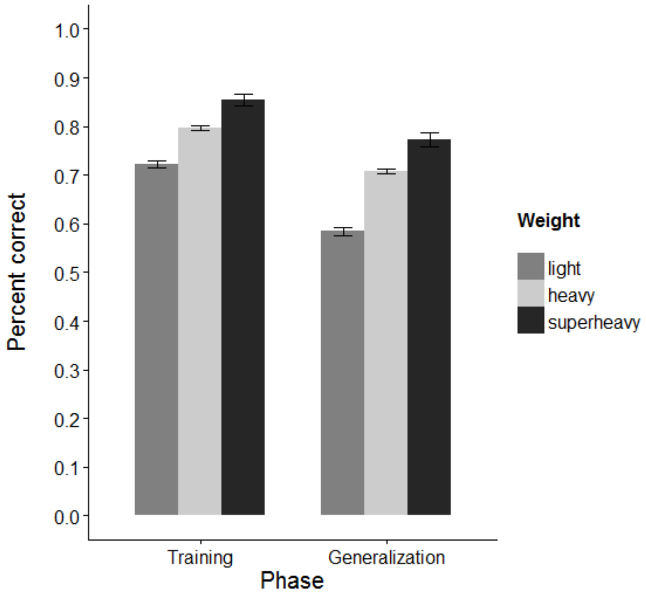


Figure A7: Correct responses by phase (training or generalization) and weight. Error bars represent standard error.

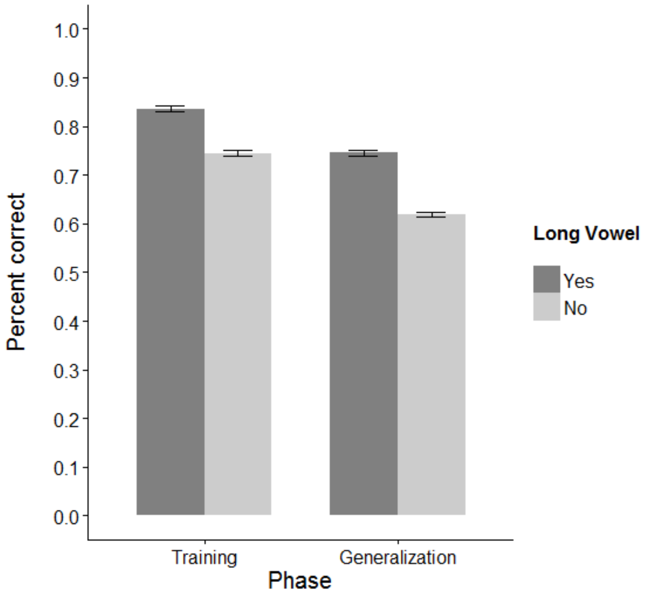


Figure A8: Correct responses by phase (training or generalization) and long vowel. Error bars represent standard error.

Order Effects

In order to check for order effects, an additional factor 'Correct_Syll_Order' was added to the statistical models in Tables 3 and 5. This factor encodes whether the correct choice was presented as the first or second item. Table A1 replicates the statistical model in Table 3 with the inclusion of 'Correct_Syll_Order'. The formula for this model is Correct ~ Constraint_Name + Day + Correct_Syll_Order + (1 | Worker_ID) + (1 | Item_Number). Table A2 replicates the statistical model in Table 5 with the inclusion of 'Correct_Syll_Order'. The formula for this model is Correct ~ Stress_Position + Weight + Long_Vowel + Day + Correct_Syll_Order + (1 | Worker_ID) + (1 | Item_Number).

	Coefficient	*SE*	*z*	*p*	
(Intercept)	0.920	0.155	5.920	< 0.001	***
EndRule	0.360	0.189	1.904	0.057	.
EndRule & WSP	0.632	0.198	3.196	0.001	**
NonFin	-0.438	0.186	-2.348	0.019	*
NonFin & EndRule	0.536	0.185	2.897	0.004	**
NonFin & EndRule & WSP	0.615	0.192	3.208	0.001	**
NonFin & WSP	0.374	0.187	1.998	0.046	*
Trochaic & A-L	-0.167	0.184	-0.906	0.365	
Trochaic & A-L & EndRule	0.056	0.189	0.297	0.767	
Trochaic & A-L & EndRule & Nonfin	0.202	0.191	1.055	0.291	
Trochaic & A-L & NonFin	-0.318	0.178	-1.782	0.075	.
WSP	0.032	0.182	0.177	0.859	
WSPμμμ	0.188	0.186	1.009	0.313	
day2	-0.248	0.125	-1.984	0.047	*
day3	-0.237	0.123	-1.928	0.054	.
day4	-0.270	0.085	-3.186	0.001	**
correct_syll_order2	-0.279	0.042	-6.652	< 0.001	***

Table A1: Constraints, Experiment Day and Correct Syllable Order with coefficient, standard error, z values and p values

	Coefficient	*SE*	*z*	*p*	
(Intercept)	1.111	0.144	7.698	< 0.001	***
stress_posINITIAL	-0.565	0.102	-5.543	< 0.001	***
stress_posFINAL_CVV	-0.501	0.157	-3.184	0.001	**
stress_posFINAL_µµµ	-0.248	0.201	-1.231	0.218	
long_vowel1	0.220	0.108	2.048	0.041	*
weightheavy	0.063	0.112	0.559	0.576	
day2	-0.117	0.115	-1.015	0.310	
day3	-0.201	0.118	-1.706	0.088	.
day4	-0.220	0.082	-2.683	0.007	**
correct_syll_order2	-0.279	0.042	-6.663	< 0.001	***

Table A2: Stress Position, Weight, Long Vowel, Day and Correct Syllable Order: coefficient, standard error, z and p values

In both models, participants' accuracy was significantly greater when the correct choice was presented first. This indicates that participants may have had a bias towards selecting the first option. However, the model additionally demonstrates that including the factor 'correct_syll_order' does not affect the conclusions in Tables 3 and 5. Effect sizes and p-values for all other factors remain the same within two decimal places.

Word Length

In order to check whether word length had a significant effect on participants' performance, an additional factor 'Num_Sylls' was added to the statistical models in Tables 3 and 5. This factor is a continuous variable which encodes how long each word is, in terms of number of syllables. Table A3 replicates the statistical model in Table 3 with the inclusion of 'Num_Sylls'. The formula for this model is Correct ~ Constraint_Name + Day + Num_Sylls + (1 | Worker_ID) + (1 | Item_Number). Table A4 replicates the statistical model in Table 5 with the inclusion of 'Num_Sylls'. The formula for this model is Correct ~ Stress_Position + Weight + Long_Vowel + Day + Num_Sylls + (1 | Worker_ID) + (1 | Item_Number).

	Coefficient	*SE*	*z*	*p*	
(Intercept)	0.947	0.239	3.969	< 0.001	***
EndRule	-0.095	0.163	-0.583	0.560	
EndRule & WSP	0.050	0.195	0.255	0.799	
NonFin	-0.618	0.162	-3.823	< 0.001	***
NonFin & EndRule	-0.138	0.160	-0.864	0.388	
NonFin & EndRule & WSP	0.254	0.187	1.355	0.175	
NonFin & WSP	-0.192	0.152	-1.257	0.209	
Trochaic & AFL	-0.632	0.149	-4.238	< 0.001	***
Trochaic & AFL & EndRule	-0.501	0.183	-2.741	0.006	**
Trochaic & AFL & Endrule & Nonfin	-0.397	0.183	-2.165	0.030	*
Trochaic & AFL & NonFin	-1.126	0.148	-7.621	0.000	***
WSP	-0.251	0.154	-1.635	0.102	
WSPμμμ	0.284	0.160	1.774	0.076	.
day2	0.133	0.078	1.707	0.088	.
day3	0.222	0.079	2.819	0.005	**
day4	0.231	0.062	3.743	< 0.001	***
num_sylls	-0.022	0.074	-0.301	0.763	

Table A3: Constraints, Experiment Day and Number of Syllables with coefficient, standard error, z values and p values

	Coefficient	*SE*	*z*	*p*	
(Intercept)	0.715	0.265	2.698	0.007	**
stress_posINITIAL	-0.485	0.116	-4.188	< 0.001	***
stress_posFINAL_CVV	-0.200	0.138	-1.452	0.147	
stress_posFINAL_μμμ	0.312	0.170	1.832	0.067	.
long_vowel1	0.373	0.086	4.344	0.000	***
weightheavy	0.206	0.090	2.276	0.023	*
day2	0.132	0.078	1.684	0.092	.
day3	0.224	0.079	2.853	0.004	**
day4	0.229	0.062	3.718	< 0.001	***
num_sylls	-0.078	0.060	-1.295	0.195	

Table A4: Stress Position, Weight, Long Vowel, Day and Number of Syllables: coefficient, standard error, z and p values

The above models demonstrate that word length is not a significant factor affecting participants' performance. This holds true both assuming that participants are acquiring OT constraint rankings, as well as assuming that participants are acquiring surface level stress patterns. However, it is possible that these results may obscure a potential effect in the case that word length interacts significantly with syllable structure. In order to rule out this possibility, an additional statistical model was run. Table A5 replicates the statistical model in Table 5 with the inclusion of 'Stress_Position * Num_Sylls'. The formula for this model is Correct ~ Stress_Position * Num_Sylls + Weight + Long_Vowel + Day + (1 | Worker_ID) + (1 | Item_Number). It was not possible to test for this interaction in the OT constraint model because this model failed to converge.

	Coefficient	*SE*	*z*	*p*	
(Intercept)	0.351	0.413	0.850	0.396	
stress_posINITIAL	0.371	0.494	0.750	0.453	
stress_posFINAL_CVV	0.457	0.663	0.688	0.491	
stress_posFINAL_μμμ	-0.015	0.762	-0.019	0.985	
num_sylls	0.159	0.102	1.568	0.117	
long_vowel1	0.198	0.106	1.863	0.062	.
weightheavy	0.106	0.120	0.888	0.375	
day2	-0.093	0.115	-0.812	0.417	
day3	-0.197	0.119	-1.651	0.099	.
day4	-0.205	0.083	-2.482	0.013	*
stress_posINITIAL: num_sylls	-0.296	0.165	-1.795	0.173	
stress_posFINAL_CVV num_sylls	-0.299	0.223	-1.337	0.181	
stress_posFINAL_μμμ: num_sylls	-0.024	0.229	-0.104	0.917	

Table A5: Stress Position * Number of Syllables, Weight, Long Vowel, and Day: coefficient, standard error, z and p values

Onsetless Syllables

Onset consonants do not affect stress in most weight-sensitive stress systems. However, rare cases of onset-sensitive stress have been observed (Gordon, 2005). Given that a number of stimuli in the current experiment contain onsetless initial syllables, it is therefore appropriate to check whether participants' performance on these words was significantly different. In order to check whether onsetless syllables had a significant effect on participants' performance, an additional factor 'Onsetless' was added to the statistical models in Tables 3 and 5. This factor is a continuous variable which encodes how long each word is, in terms of number of syllables. Table A6 replicates the statistical model in Table 3 with the inclusion of 'Onsetless'. The formula for this model is Correct ~ Constraint_Name + Day + Onsetless + (1 | Worker_ID) + (1 | Item_Number). Table A7 replicates the statistical model in Table 5 with the inclusion of 'Onsetless'. The formula for this model is Correct ~ Stress_Position + Weight + Long_Vowel + Day + Onsetless + (1 | Worker_ID) + (1 | Item_Number).

	Coefficient	*SE*	*z*	*p*	
(Intercept)	0.775	0.153	5.050	< 0.001	***
EndRule	0.393	0.195	2.013	0.044	*
EndRule & WSP	0.659	0.203	3.248	0.001	**
NonFin	-0.418	0.187	-2.232	0.026	*
NonFin & EndRule	0.568	0.187	3.042	0.002	**
NonFin & EndRule & WSP	0.633	0.194	3.263	0.001	**
NonFin & WSP	0.368	0.187	1.969	0.049	*
Trochaic & AFL	-0.166	0.184	-0.904	0.366	
Trochaic & AFL & EndRule	0.063	0.189	0.332	0.740	
Trochaic & AFL & Endrule & Nonfin	0.228	0.195	1.168	0.243	
Trochaic & AFL & NonFin	-0.310	0.178	-1.741	0.082	.
WSP	0.035	0.182	0.190	0.850	
WSPμμμ	0.207	0.186	1.115	0.265	
day2	-0.243	0.125	-1.939	0.052	.
day3	-0.239	0.123	-1.950	0.051	.
day4	-0.267	0.085	-3.151	0.002	**
onsetlessyes	-0.071	0.099	-0.710	0.478	

Table A6: Constraints, Experiment Day and Onsetless Syllables with coefficient, standard error, z values and p values

	Coefficient	*SE*	*z*	*p*	
(Intercept)	0.975	0.144	6.748	< 0.001	***
stress_posINITIAL	-0.569	0.103	-5.526	< 0.001	***
stress_posFINAL_CVV	-0.509	0.160	-3.183	0.001	**
stress_posFINAL_μμμ	-0.239	0.201	-1.188	0.235	
long_vowel1	0.218	0.108	2.024	0.043	*
weightheavy	0.064	0.112	0.572	0.568	
day2	-0.117	0.115	-1.015	0.310	
day3	-0.203	0.118	-1.723	0.085	.
day4	-0.218	0.082	-2.666	0.008	**
onsetlessyes	-0.035	0.096	-0.363	0.717	

Table A7: Stress Position, Weight, Long Vowel, Day and Onsetless Syllables: coefficient, standard error, z and p values

The above models demonstrate that participants' performance is not significantly affected by the presence of onsetless syllables. This holds true for both the OT constraint model and the surface stress pattern model.

APPENDIX B: STIMULI

Day 1: Constraint group WSP

Familiarization Phase

'taa.wa	HL
'in.si	HL
ga.'lam.bu	LHL
di.'raa.si	LHL
wa.gi.'bat.li	LLHL

Generalization Phase

'faa.tu	HL
'ʃat.mu	HL
ma.'gaa.lu	LHL
bu.'lan.di	LHL
di.na.'mii.ki	LLHL

Generalization Phase (Day 4)

'ruk.ni	HL
'ʃuu.fi	HL
a.'saa.mi	LHL
ka.'lam.na	LHL
ku.tu.'muu.tu	LLHL

Day 1: Constraint group WSPμμμ

Familiarization Phase		*Generalization Phase*	
mus.tan.za.'maat	HHLS	ma.tin.ʃi.'gilʃ	LHLS
it.fat.'wint	HHS	mab.tin.'zilʃ	HHS
baʃ.'niin	HS	sa.'fruut	HS
ma.ba.'namʃ	LLS	wa.fa.'daan	LLS
za.la.'laan	LLS	sa.ka.'lans	LLS

Generalization Phase (Day 4)

taʃ.ta.ra.'waan	HHLS
it.nam.'fizt	HHS
tan.'ziim	HS
ta.ra.'biiz	LLS
fu.ru.'sint	LLS

Day 1: Constraint group NONFIN

Familiarization Phase

'rik.bit	HH
'maa.ziz	HH
'ba.nit	LH
'za.mat	LH
'ki.niz	LH

Generalization Phase

'ʃuf.lak	HH
'ʃaa.win	HH
'ma.san	LH
'wa.jam	LH
'ga.bad	LH

Generalization Phase (Day 4)

'bad.ris	HH
'gaa.lis	HH
'fi.rig	LH
'ʃa.naf	LH
'da.mak	LH

Day 1: Constraint group ENDRULE-R

Familiarization Phase		*Generalization Phase*	
is.'tab.da	HHL	um.'baa.ʃi	HHL
as.'wan.li	HHL	ig.'maa.li	HHL
ʃam.'bat.li	HHL	bar.'kit.li	HHL
gib.'tuu.li	HHL	gin.'sii.tu	HHL
da.waʃ.'naa.ku	LHHL	mi.nas.'baa.lu	LHHL

Generalization Phase (Day 4)

it.'naa.da	HHL
ab.'rii.mi	HHL
mis.'tas.ni	HHL
tag.'rii.di	HHL
za.man.'kaa.wi	LHHL

Day 1: Constraint group WSP-CVV

Familiarization Phase

fak.'raa	HH
lab.'saa	HH
ʃuf.'tii	HH
bi.tin.'saa	LHH
ba.la.'dii	LLH

Generalization Phase

taf.'taa	HH
jiʃ.'fii	HH
ʃin.'waa	HH
ʃa.tam.'tii	LHH
da.fa.'nuu	LLH

Generalization Phase (Day 4)

gib.'naa	HH
nas.'jaa	HH
gam.'bii	HH
fa.raʃ.'naa	LHH
mu.ba.'laa	LLH

Day 2: Constraint group ENDRULE-R & WSP

Familiarization Phase

mis.til.'mii.nu	HHHL
an.ti.'kan.ja	HLHL
iʃ.ti.'rii.li	HLHL
kis.ti.'naa.wi	HLHL
bi.jif.tik.'ruu.ni	LHLHL

Generalization Phase

ban.tis.'ban.ja	HHHL
jus.ta.'fan.di	HLHL
tis.ta.'fii.du	HLHL
muk.li.'maa.ni	HLHL
mi.bas.ba.'saa.ti	LHLHL

Generalization Phase (Day 4)

is.tif.'zaa.zi	HHHL
in.ti.'waa.zi	HLHL
ik.ti.'bii.li	HLHL
taʃ.ri.'faa.ti	HLHL
bi.jiʃ.ti.'kii.li	LHLHL

Day 2: Constraint group TROCHAIC & ALIGN-L

Familiarization Phase		*Generalization Phase*	
'za.ki	LL	'gi.za	LL
'bi.la	LL	'ka.ba	LL
'ga.tu	LL	'sa.da	LL
'lu.ta.ri	LLL	'di.na.mu	LLL
'za.ga.li	LLL	'nu.ka.ti	LLL

Generalization Phase (Day 4)

'du.ga	LL
'ʃi.fa	LL
'dʒi.li	LL
'ga.ma.li	LLL
'sa.ka.ni	LLL

Day 2: Constraint group NonFin & WSP

Familiarization Phase

ti.'nas.bak	LHH
mi.'kam.fit	LHH
mi.'zam.bin	LHH
da.'faa.tir	LHH
li.'saa.nik	LHH

Generalization Phase

bi.'nis.nid	LHH
di.'ras.tak	LHH
la.'zaz.tik	LHH
la.'daa.jin	LHH
wi.'laa.dak	LHH

Generalization Phase (Day 4)

mi.'tam.fis	LHH
mi.'faz.lik	LHH
ta.'san.sun	LHH
mi.'gaa.nis	LHH
fi.'luu.sak	LHH

Day 2: Constraint group NonFin & EndRule-R

Familiarization Phase

an.'faa.sik	HHH
saj.'bin.lik	HHH
is.'tad.sim	HHH
gi.lig.'naa.jit	LHHH
sa.la.'mit.kum	LLHH

Generalization Phase

mat.'saa.fin	HHH
it.'ʃan.kam	HHH
mit.'bas.tif	HHH
bi.tit.'naa.win	LHHH
ta.ta.'naa.sab	LLHH

Generalization Phase (Day 4)

it.'kas.kis	HHH
bit.'zaa.kin	HHH
mis.'taʃ.kil	HHH
ta.man.'taa.ʃan	LHHH
mu.ta.'faa.wit	LLHH

Day 3: Constraint group Trochaic & Align-L & EndRule-R & Parse-Ft

Familiarization Phase

mit.fat.'fi.ta	HHLL
aw.fan.'li.na	HHLL
a.gib.'lu.ku	LHLL
mu.zak.'ri.tu	LHLL
bi.tiʃ.'ra.bu	LHLL

Generalization Phase

hat.ban.'li.na	HHLL
mit.ban.'ʃi.ma	HHLL
sa.ban.'si.gi	LHLL
ba.rik.'ti.lu	LHLL
mi.dam.'bi.ka	LHLL

Generalization Phase (Day 4)

is.tag.'fi.ru	HHLL
mus.tam.'ti.ka	HHLL
ba.naf.'si.gi	LHLL
bi.tim.'si.ki	LHLL
ti.ʃuf.'li.na	LHLL

Day 3: Constraint group TROCHAIC & ALIGN-L & NONFIN

Familiarization Phase

'ka.ti.fan	LLH
'ʃa.ra.dit	LLH
'na.za.fit	LLH
'ma.za.lan	LLH
'ma.ba.sak	LLH

Generalization Phase

'ta.ra.lan	LLH
'a.ba.dan	LLH
'ki.ra.win	LLH
'na.ʃa.zit	LLH
'ba.la.dak	LLH

Generalization Phase (Day 4)

'ka.ta.bit	LLH
'ka.ʃa.fit	LLH
'fa.ʃa.lit	LLH
'da.ra.sit	LLH
'ra.ka.sit	LLH

Day 3: Constraint group NONFIN & ENDRULE-R & WSP

Familiarization Phase		*Generalization Phase*	
maʃ.ta.'rak.tiʃ	HLHH	is.ta.'ʃan.tak	HLHH
iʃ.ta.'gan.tum	HLHH	is.bi.'laa.jit	HLHH
tim.bi.'sii.lik	HLHH	mus.ta.'waa.kum	HLHH
mis.ti.'baa.rik	HLHH	jin.gu.'duu.lak	HLHH
in.ti.'zaa.rik	HLHH	il.ma.'baa.lig	HLHH

Generalization Phase (Day 4)

is.ti.'kan.tik	HLHH
if.ta.'kan.tak	HLHH
baʃ.ti.'kii.lak	HLHH
mis.ti'kaa.min	HLHH
ib.ti.'daa.kan	HLHH

Day 3: Constraint group ALIGN-L & ENDRULE-R & NONFIN

Familiarization Phase		*Generalization Phase*	
ha.niʃ.'ta.gal	LHLH	bi.tit.'ki.sif	LHLH
ka.tab.'ti.lik	LHLH	ha.nig.'si.bik	LHLH
bi.nak.'ta.sin	LHLH	ji.gib.'lu.kum	LHLH
ma.ʃuf.'tu.kum	LHLH	bi.tak.'ta.mid	LHLH
mus.tak.'ba.lan	LHLH	ʃa.tam.'tu.kum	LHLH

Generalization Phase (Day 4)

mu.kal.'mi.tak	LHLH
bi.tit.'li.bis	LHLH
ha.tit.'ki.ʃif	LHLH
bi.jan.'da.mig	LHLH
ta.lab.'tu.kum	LHLH

REFERENCES

Altmann, H. (2006). *The perception and production of second language stress: A cross-linguistic experimental study*. University of Delaware.

Alvord, S. M. (2003). The psychological unreality of quantity sensitivity in Spanish: experimental evidence. *Southwest Journal of Linguistics*, *22*(2), 1–13.

Anani, M. (1989). Incorrect stress placement in the case of Arab learners of English. *IRAL-International Review of Applied Linguistics in Language Teaching*, *27*(1), 15–22.

Archibald, J. (1993). *Language Learnability and L2 Phonology*. Dordrecht: Kluwer Academic Publishers.

Archibald, J. (1997). The Acquisition of English Stress by Speakers of Nonaccentual Languages: Lexical Storage versus Computation of Stress. Linguistics 35, 1, 167-181.

Archibald, J. (2009). The acquisition of English stress by speakers of nonaccentual languages: lexical storage versus computation of stress. *Linguistics*, *35*(1), 167–182.

Baker, R. G., & Smith, P. T. (1976). A psycholinguistic study of English stress assignment rules. *Language and Speech*, *19*(1), 9–27.

Baptista, B.O. (1989). Strategies for the Prediction of English Word Stress. *International Review of Applied Linguistics* 27, 1, Feb,,1-14.

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Broselow, E. (1976). The phonology of Egyptian Arabic.

Broselow, E. (1979). Cairene Arabic syllable structure. *Linguistic Analysis* 5, 345-382.

Broselow, E., M. Huffman, S. Chen & R. Hsieh (1995). The timing structure of CVVC syllables. In M. Eid (ed.) *Perspectives on Arabic linguistics VII.* Amsterdam: Benjamins. 119-140.

Broselow, E., Chen, S. I., & Huffman, M. (1997). Syllable weight: convergence of phonology and phonetics. *Phonology*, *14*(01), 47-82.

Burzio, L. (1994). *Principles of English stress*. Cambridge [England] ; New York: Cambridge University Press.

Carpenter, A. C. (2005). Acquisition of a natural vs. unnatural stress system. In *Proceedings of the 29th annual Boston University Conference on Language Development* (pp. 134–43).

Carpenter, A. C. (2010). A naturalness bias in learning stress. *Phonology*, *27*(3), 345–392.

Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English*. New York: Harper & Row.

Christiansen, M. H. (2000). Using artificial language learning to study language evolution: Exploring the emergence of word order universals. In *The evolution of language: 3rd international conference* (pp. 45–48).

Clopper, C. G. (2002). Frequency of stress patterns in English: A computational analysis. *IULC Working Papers Online*, *2*(2).

Crowhurst, M. J. (1996). An optimal alternative to conflation. *Phonology*, *13*(3), 409–424.

Culbertson, J. (2012). Typological Universals as Reflections of Biased Learning: Evidence from Artificial Language Learning. *Language and Linguistics Compass*, *6*(5), 310–329.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, *122*(3), 306–329.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech & Language*, *2*(3–4), 133–142.

Davidson, L., & Shaw, J. A. (2012). Sources of illusion in consonant cluster perception. *Journal of Phonetics*, *40*(2), 234-248.

Davidson, L. (2010). Phonetic bases of similarities in cross-language production: Evidence from English and Catalan. *Journal of Phonetics*, *38*(2), 272-288.

Davis, S. M., & Kelly, M. H. (1997). Knowledge of the English Noun–Verb Stress Difference by Native and Nonnative Speakers. *Journal of Memory and Language*, *36*(3), 445–460.

De Lacy, P. V. (2002). *The formal expression of markedness*. University of Massachusetts Amherst.

Domahs, U., Plag, I., & Carroll, R. (2014). Word stress assignment in German, English and Dutch: Quantity-sensitivity and extrametricality revisited. *The Journal of Comparative Germanic Linguistics*, *17*(1), 59–96.

Dupoux, E., & Peperkamp, S. (2002). Fossil markers of language development: phonological “deafnesses” in adult speech processing. *Phonetics, Phonology, and Cognition*, 168–190.

Dupoux, E., Peperkamp, S., & Sebastián-Gallés, N. (2001). A robust method to study stress “deafness.” *The Journal of the Acoustical Society of America*, *110*(3), 1606–1618.

Erdmann, P. H. (1973). Patterns of stress-transfer in English and German. *International Review ofApplied Linguistics*, 11, 299-241.

Fedzechkina, M., Jaeger, T., & Newport, E. (2011). Functional biases in language learning: Evidence from word order and case-marking interaction. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 318–323).

Gouskova, M. (2003). *Deriving economy: syncope in Optimality Theory*. University of Massachusetts Amherst.

Green, T., & Kenstowicz, M. (1995). The lapse constraint.

Guest, D. J., Dell, G. S., & Cole, J. S. (2000). Violable Constraints in Language Production: Testing the Transitivity Assumption of Optimality Theory. *Journal of Memory and Language*, *42*(2), 272–299.

Guion, S. G. (2005). KNOWLEDGE OF ENGLISH WORD STRESS PATTERNS IN EARLY AND LATE KOREAN-ENGLISH BILINGUALS. *Studies in Second Language Acquisition*, *27*(4), 503–533.

Guion, S. G., Clark, J. J., Harada, T., & Wayland, R. P. (2003). Factors Affecting Stress Placement for English Nonwords include Syllabic Structure, Lexical Class, and Stress Patterns of Phonologically Similar Words. *Language and Speech*, *46*(4), 403–426.

Guion, S. G., Harada, T., & Clark, J. J. (2004). Early and late Spanish–English bilinguals' acquisition of English word stress patterns. *Bilingualism: Language and Cognition*, *7*(3), 207–226.

Halle, Morris and Jean-Roger Vergnaud (1987) *An essay on stress*. Cambridge, MA.: MIT Press.

Hayes, B. (1982). Extrametricality and English Stress. *Linguistic Inquiry*, *13*(2), 227–276.

Hayes, B. (1995). *Metrical stress theory: Principles and case studies*. University of Chicago Press.

Hewitt, M. (1994). Deconstructing foot binarity in Koniag Alutiiq. *Ms, University of British Columbia. Available as ROA-12 from the Rutgers Optimality Archive*.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*(2), 151–195.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*(4), 548–567.

Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Infants' Preference for the Predominant Stress Patterns of English Words. *Child Development*, *64*(3), 675–687.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, *39*(3), 159–207.

Kager, R. (2001) "Rhythmic Directionality by Positional Licensing", handout of a presentation given at HILP-5, University of Potsdam. ROA-514.

Kager, R. W. J. (2005). Rhythmic licensing theory: an extended typology. In*Proceedings of the third international conference on phonology* (pp. 5-31). Seoul National University.

Kilany, Hanaa, et al. *Egyptian Colloquial Arabic Lexicon LDC99L22.* Web Download. Philadelphia: Linguistic Data Consortium, 1997.

Lassiter, D. (2014) Submiterator. Retrieved from http://github.com/danlassiter/Submiterator

Lord, G. E. (2001). *The second language acquisition of Spanish stress: Derivational, analogical or lexical?* (Ph.D.). The Pennsylvania State University, United States -- Pennsylvania.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*(1), 11–28.

Mairs, J. L. (1989). Stress assignment in interlanguage phonology: An analysis of the stress system of Spanish speakers learning English. *Linguistic perspectives on second language acquisition*, 260-283.

McCarthy, J. (1979). On Stress and Syllabification. *Linguistic Inquiry*, *10*(3), 443–465.

McCarthy, J. (2003). OT constraints are categorical. *Phonology*, *20*(1), 75–138.

McCarthy, J. J., & Prince, A. (1993). *Generalized alignment*. Springer.

Moreton, E. (2008). Analytic bias and phonological typology. *Phonology*, *25*(1), 83.

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language learning, 49*(1), 285-310.

Pater, J. (1997). Metrical parameter missetting in second language acquisition. *Language acquisition and language disorders*, *16*, 235-262.

Peperkamp, S. (2004). Lexical Exceptions in Stress Systems: Arguments from Early Language Acquisition and Adult Speech Perception. *Language*, *80*(1), 98–126.

Peperkamp, S., & Dupoux, E. (2002). A typological study of stress “deafness.” *Laboratory Phonology*, *7*, 203–240.

Peperkamp, S., Vendelin, I., & Dupoux, E. (2010). Perception of predictable stress: A cross-linguistic investigation. *Journal of Phonetics*, *38*(3), 422–430.

Prince, A. (1983). Relating to the grid. *Linguistic Inquiry*, 19–100.

Prince, A. (1990). Quantitative consequences of rhythmic organization. *CLS*, *26*(2), 355–398.

Prince, A., & Smolensky, P. (1993/2002). *Optimality Theory: Constraint Interaction in Generative Grammar* (Technical).

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*(6), 855–863.

Salsignac, J. (1998). Une étude sur la perception de l’accent primaire de langues étrangères. *Letras de Hoje* 33, 4, 81-107.

Schembri, T., Johnson, M., & Demuth, K. (2016) Longi: A Simple Automated System for Conducting Longitudinal

Studies on Amazon Mechanical Turk. Manuscript in preparation.

Selkirk, E. O. (1980). The role of prosodic categories in English word stress. *Linguistic Inquiry* 11, 563-605.

Sereno, J. A. (1986). Stress pattern differentiation of form class in English. *The Journal of the Acoustical Society of America*, *79*(S1), S36–S36.

Shaw, J. A., & Davidson, L. (2011). Perceptual similarity in input–output mappings: A computational/experimental study of non-native speech production. *Lingua*, *121*(8), 1344-1358.

St Clair, M. C., Monaghan, P., & Ramscar, M. (2009). Relationships between language structure and language learning: The suffixing preference and grammatical categorization. *Cognitive Science*, *33*(7), 1317–1329.

Taylor, P., Black, A. W., & Caley, R. (1998). The Architecture Of The Festival Speech Synthesis System. In *IN THE THIRD ESCA WORKSHOP IN SPEECH SYNTHESIS* (pp. 147–151).

Tesar, B. (1997). An iterative strategy for learning metrical stress in Optimality Theory. In *Proceedings of the 21st Annual Boston University Conference on Language Development* (pp. 615–626).

Tremblay, A., & Owens, N. (2010). The role of acoustic cues in the development of (non-) target-like second-language prosodic representations. *The Canadian Journal of Linguistics / La Revue Canadienne de Linguistique*, *55*(1), 85–114.

Wayland, R., Landfair, D., Li, B., & Guion, S. G. (2006). Native Thai Speakers’ Acquisition of English Word Stress Patterns. *Journal of Psycholinguistic Research*, *35*(3), 285–304.

Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*, *30*(5), 945–982.

Chapter 4: Is Less Really More? Answers from L2 Stress Acquisition

ABSTRACT

Young children acquire language without explicit tuition. Conversely, adult learners often experience significant difficulties in second language acquisition, despite superior cognitive capabilities. The "less is more" hypothesis holds that children's limited working memory capacity facilitates acquisition of their native language. Hence, according to this hypothesis, adults' increased cognitive capabilities actively hinders them while learning a second language. A body of research has attempted to simulate limited working memory in adults by limiting the input initially available to them. If the "less is more" hypothesis is correct, this should improve performance when compared to participants who are immediately presented with stimuli of all levels of complexity. The current study tests this hypothesis with adult participants acquiring the Cairene Arabic stress system. Participants were presented with randomly intermixed stimuli; items of all levels of complexity were presented at any point in the experiment. The results were compared to an earlier experiment in which stimuli of gradually increasing complexity were used; all other aspects of the two experiments were identical. Participants in the random experiment outperformed those in the earlier experiment; additionally, performance was improved on all metrics measuring knowledge of the underlying prosodic structure of the language. These results demonstrate that, at least for the narrow domain of word stress acquisition, the 'less is more' hypothesis does not hold.

1. Introduction

Language learning is exceptionally difficult for adult learners, who rarely achieve native speaker competence in a language acquired in adulthood. Yet children are able to easily acquire their native language, and ultimately achieve full competence. This is in spite of adults' superior cognitive capabilities, which allow adults to learn most non-linguistic skills faster than children. Thus, one of the defining challenges for language acquisition research is to explain why it is that children's language acquisition skills are so superior. A number of competing hypotheses have been advanced to explain these observations. The “less is more”, or “starting small”, hypothesis posits that children’s limited working memory capacity is helpful in learning language. Because of this limited capacity, children perceive and store only small sections of speech. This allows them to decompose the speech stream into smaller linguistic units, such as morphemes and phonemes (Newport, 1990; Elman, 1993; Gibbs, 2004). For example, Newport (1990) reports that adult learners of American Sign Language (ASL) use “frozen” unanalyzed structures which are reproduced as a whole, without any understanding of their internal morphological structure. She argues that this is because adult learners are able to store entire stretches of speech, such as whole sentences. Breaking up these large sections of speech into their smallest components is considered a more difficult task, because there are a greater number of combinatorial possibilities to consider. Therefore, according to this hypothesis, adults’ inferior language acquisition skills are, at least in part, due to their superior cognitive and information processing capabilities. Crucially, this provides a testable hypothesis: if this is true, then acquisition in adults can be improved by either simplifying or limiting the input initially provided (Antoniou, Ettlinger, & Wong, 2016; Arnon & Ramscar, 2012; Chin, 2009; Kersten & Earles, 2001; Lai & Poletiek, 2010, 2011, 2013) or by simulating cognitive limitations (Cochran, McDonald, & Parault, 1999; Ludden & Gupta, 2000); this is done by asking participants to carry out cognitively challenging tasks while involved with the language learning task. These issues are explored in further detail in the section below.

1.1 Testing the “Less is More” hypothesis

Kersten & Eales (2001) taught an artificial language encoding object, path and manner of motion to adult participants under two conditions. In the first condition, participants initially heard individual words, and only later progressed to full sentences; in the second condition, participants were presented with sentences throughout. Participants whose input was initially limited performed better on both word learning and morphology than those who were immediately exposed to the full complexity of the language. This result was argued to support Newport's 'Less is More' hypothesis. Chin (2009) carried out a similar experiment, using French rather than an artificial language: adult participants with no prior experience of a Romance language were taught French active and reflexive verb forms. In one condition, participants initially heard short phrases, such as 'la voiture' (the car), and gradually progressed to full sentences; in another, participants were exposed to full sentences throughout. Participants who began the experiment with limited input scored higher on grammar tasks than those who were exposed to full sentences from the start. These results are consistent with Newport's hypothesis. The above experiments test the 'Less is More' hypothesis by limiting the input based on stimulus length.

Another set of experiments apply this idea to grammatical complexity rather than stimulus length. In these experiments, one set of participants is initially presented with grammatically simple input, which gradually increases in complexity; in a second condition, participants are presented with the full complexity of the system throughout the experiment. Syntactic recursion is ideal for this purpose, as complexity can be increased indefinitely: a string can contain no recursion, such as "Mary is beautiful"; a single level of recursion, such as "[John thinks [Mary is beautiful]]"; two levels of recursion, such as "[Peter knows that [John thinks [Mary is beautiful]]]", and so on. Grammatical complexity can be gradually and indefinitely increased in this manner. Participants were taught a recursive system under two conditions: in the first condition, the complexity of the input was initially limited; in the second condition, stimuli were presented in random order, with items of all levels of complexity presented in any order. Participants in the limited condition performed significantly better than those in the random condition; additionally, participants in the random condition were unable to acquire the recursive rule (Lai & Poletiek, 2010, 2011, 2013). These experiments demonstrate that the 'less is more' hypothesis is supported when input is limited based on grammatical complexity, as well as stimulus length. Similarly, Antoniou et al. (2016) apply this principle to morphophonological, rather than syntactical, grammatical complexity. Participants were taught a system with both simple and complex morphophonological components: the simple rule involved concatenating noun stems with prefixes or suffixes; the complex rule required participants to learn three linguistic processes simultaneously: affix concatenation, vowel harmony affecting both stem and affix, and vowel reduction affecting the noun stem only. Participants were assigned to one of two conditions: in the simple-first condition, participants were presented with simple input, followed by complex input; in the complex-first condition, the order of presentation was reversed. Participants in the simple-first condition performed significantly better on both simple and complex components of the grammar. These studies demonstrated that the 'Less is More' hypothesis can be applied to grammatical complexity as well as stimulus length. The results support Newport's hypothesis, showing that participants performed better when grammatically simple forms were introduced before grammatically complex forms.

The above experiments all demonstrate that the 'less is more' hypothesis can be tested by providing participants with initially limited input. Another way of testing this hypothesis is to simulate cognitive limitations in adults; if the 'less is more' hypothesis is correct, this should lead to improved performance in language acquisition. Cochran et al. (1999) asked one set of participants to carry out an unrelated task, designed to increase cognitive load, while learning ASL signs. They found that participants who acquired the language under conditions involving additional processing load produced certain signs more accurately than those who acquired the language under normal conditions. They concluded that the resultant processing limitations enabled participants to decompose signs into their constituent morphemes, while participants who acquired the signs under normal conditions produced unanalysed signs.

All the above studies appear to demonstrate that the "less is more" hypothesis is correct. However, a competing body of research demonstrates the opposite result (Arnon & Ramscar, 2012; Ludden & Gupta, 2000; Rohde & Plaut, 1999; Siegelman & Arnon, 2015). Arnon & Ramscar (2012) taught participants an artificial language encoding grammatical gender through the use of articles. The methodology was similar to that of Kersten & Eales (2001) and Chin (2009): one group of adult participants were first presented with article-noun sequences in full sentences, and only later encountered bare nouns without articles; the other group was first exposed to isolated nouns, and then to full sentences. Crucially, both sets of participants were exposed to exactly the same input, but in different orders. Participants who were first exposed to full sentences performed better on both word learning and acquisition of grammatical gender; in addition, they were quicker to produce correct responses in a production task. This illustrated a case in which the 'less is more' hypothesis did not hold for input limited by stimulus length. Ludden & Gupta (2000) used methodology similar to that in Cochran et al. (1999), in which one set of participants performed linguistic tasks under cognitive load, in order to simulate cognitive limitations. In the Cognitive Load condition, participants were asked to draw a picture while listening to stimuli; in the No Load condition, participants listened to stimuli without carrying out an additional task. In one experiment, adult participants were given the task of segmenting words from the speech stream; in another, a pattern of simple syntactic agreement was presented. Contrary to the results in Cochran et al. (1999), the researchers found that, for both experiments, participants in the No Load condition outperformed those who acquired the system under cognitive load. These results refute the findings of Cochran et al. (1999), and demonstrate that the 'less is more' hypothesis does not always hold when memory limitations are simulated in adults.

Conway, Ellefson, & Christiansen (2003), noting these inconsistent results, set out to quantify the conditions under which the "less is more" hypothesis holds true. They carried out four artificial language experiments with adult participants, varying across two parameters. Stimuli were presented either visually, or auditorily; stimuli contained either center-embedded recursion, such as ‘the boy [the girl loves] likes the dog’, or right-branching recursion, such as, ‘the boy likes the dog [that the girl loves]’. Varying these parameters resulted in four experiments, in which participants were presented with: visual center-embedded recursive input; visual right-branching recursive input; auditory center-embedded recursive input; auditory right-branching recursive input. In each experiment, participants were assigned to one of three conditions: Starting Small, Random or Control. In the Starting Small condition, participants were trained using staged input based on the level of recursion; for example, the first block contained sentences with no recursion, while the second block contained sentences with a single level of recursion, and so forth. In the Random condition, participants received training inputs of all levels of recursion, in random order. As in Arnon & Ramscar (2012), participants in the Starting Small and Random conditions received identical input by the end of the experiment; additionally, the control group received no training. When stimuli were visually presented, participants in the starting small condition were able to acquire both types of recursive input, while those in the random condition and control group could not. However, when stimuli were aurally presented, there was no significant difference between the two participant groups. The researchers conclude that starting small aids acquisition only when input is presented visually. In support of this claim, they note that research which shows an advantage for starting small incorporates visual input, sometimes in combination with auditory input (Cochran et al., 1999; Kersten & Earles, 2001), while research which demonstrates the opposite result uses only visual input (Ludden & Gupta, 2000). However, subsequent research (Arnon & Ramscar, 2012) calls this claim into question, demonstrating an advantage for starting big while using both visual and auditory stimuli.

The “less is more” hypothesis has been primarily tested in the domain of morphosyntax. However, one study has addressed this claim in the context of prosody; notably this is the only experiment which directly compares the performance of adults and children. Kapatsinski, Olejarczuk, & Redford (2016) exposed children and adults to different intonation contours exemplifying three categories: a flat contour; a final-fall contour; an ‘M-shaped’ contour. Participants were then asked to categorize novel contours as one of these contours, or as an unknown contour. Children and adults exhibited different patterns of categorization: younger children accepted stimuli which differed greatly from the exemplars presented during training; while older children and adults rejected many more stimuli as members of a given category. The researchers concluded that children formed broader categories, while older children and adults formed narrow categories, with more stringent criteria for category membership. These results demonstrate that younger children’s categories are more underspecified than those of adults, thus supporting the “less is more” hypothesis; however, they do not causally link this underspecification to children’s superior language acquisition skills.

1.2 Original study: staged input

Schembri, Johnson, & Demuth (2016a) taught English-speaking adult participants the Cairene Arabic stress system over the course of 4 days using a staged input scheme in which stimuli were initially limited according to grammatical complexity. Stimuli were organized according to complexity based on the number of interacting constraints necessary to correctly stress each stimulus. On the first day, participants were presented with stimuli which satisfied a single, simple constraint. Stimuli presented on subsequent days required the satisfaction of multiple interacting constraints. Participants were able to acquire some aspects of the stress system, correctly stressing 64-70% of words in the generalization phase for each day. However, acquisition did not improve linearly: adjusted for other factors, participants' performance was best on the first day of the experiment, but decreased as more complex input was introduced on subsequent days.

1.3 Current study: random input

The current study aims to further investigate the "less is more" hypothesis. As in Conway (2003) and Arnon & Ramscar (2012), an experiment using staged input, in which stimuli are initially limited in grammatical complexity (Schembri et al., 2016a) is replicated using randomly organized stimuli. Crucially, the stimuli, methodology and presentation in Schembri et al. (2016a) are the same in all aspects other than order of presentation, such that both sets of participants receive identical input by the end of the experiment. This enables a direct comparison of adults' acquisition of: a) initially limited stimuli; versus b) stimuli immediately containing the full range of complexity. This kind of direct comparison has been previously limited to the domain of morphosyntax (Kersten & Eales, 2001; Chin, 2009; Arnon & Ramscar, 2012; Antoniou, Ettlinger, & Wong, 2016; Conway, 2003; Cochran, McDonald, & Parault, 1999; Lai & Poletiek, 2010, 2011, 2013; Ludden & Gupta, 2000); here it is extended for the first time to prosodic acquisition. Previous research testing the 'less is more' hypothesis has been inconclusive and contradictory; testing within the domain of prosodic acquisition may shed further light on the conditions under which "less is more".

Previous research has suggested a greater advantage for starting small when the input is more complex (Conway, 2003; Antoniou et al., 2016). The Cairene Arabic stress system has been studied extensively due to its complex, interacting constraints. The input presented to participants in Schembri et al. (2016a) captured the phonological complexity of the language as closely as possible in terms of syllable structure, word length, phonemic inventory and vocabulary, gradually introducing more complexity over a 3-day period. Conway (2003) found that the use of initially limited input is particularly advantageous when visual stimuli are included. In the current experiment, both visual and auditory stimuli are used; if Conway's hypothesis is correct, this should maximize the possibility of a 'less is more' advantage.

If participants in the staged input experiment outperform those in the random experiment, this would support the 'less is more' hypothesis. Conversely, if the reverse is true, this would demonstrate a 'starting big' advantage for adult participants learning a complex stress system; additionally, it would suggest that the use of visual stimuli does not always provoke a 'less is more' effect.

2. Material and methods

2.1 Stimuli

The stimulus words in the current experiment were identical to those in Schembri et al. (2016a). A subset of the LDC Colloquial Egyptian Arabic lexicon (Kilany et al., 1997) was used as the source, with the following sets excluded: a) one-syllable words; b) words longer than 5 syllables; c) words with multiple pronunciations; d) words of foreign origin; e) words which resembled English lexical items. The consonant inventory included only phonemes present in both American English and Cairene Arabic. Only the 3 short vowels /ɑ/, /ɪ/ and /ʊ/, and their corresponding long vowels, were included. Stimuli were produced by the Festival synthesis software (Taylor, Black, & Caley, 1998); their intensity, pitch and duration was manipulated using Praat (Boersma, 2002). The stimulus creation techniques are described in greater detail in Schembri et al. (2016a); stimuli in the current experiment were not re-synthesized or re-manipulated in any way.

Colloquial Cairene Arabic is a quantity-sensitive language with a three-way distinction between light (CV), heavy (CVC, CVV) and superheavy (CVCC, CVVC) syllables. The location of stress is entirely predictable by metrical structure, and can be determined without exception if the sequence of light (L), heavy (H) and superheavy (S) syllables is known. An Optimality Theory (OT) analysis of the Cairene Arabic stress system was carried out in Schembri et al. (2016a). The words in the corpus were organized according to the active constraints required to assign stress correctly. A total of 13 sets of OT constraint rankings, or constraint groups, was required to fully describe the stress system: NonFin; EndRule-R; WSP-CVV; WSP; WSP$_{\mu\mu\mu}$; EndRule-R + WSP; NonFin + EndRule-R; NonFin + WSP; Trochaic + Align-L; NonFin & EndRule-R & WSP; Align-L & EndRule-R & NonFin; Trochaic & Align-L & NonFin; Trochaic & Align-L & EndRule-R & Parse-Ft. For each constraint group, a set of 15 stimuli was included. A full list of stimuli used in the current experiment is included in the Appendix, organized by constraint group.

2.2 Recruitment and data collection

The experiment was created using Javascript. Submiterator (Lassiter, 2014) was used to submit completed HITs to Mechanical Turk. Longi (Schembri, Johnson, & Demuth, 2016b) was used to handle the technical issues involved in running a multi-day study. Amazon Mechanical Turk, a crowdfunding platform, was used to recruit 119 participants. All participants were required to be American English speakers with a prior approval rating of at least 95%. The experiment took place over 4 consecutive days; each daily session was completed in 20-30 minutes. Each session was posted on a regular 24-hour schedule; participants were able to complete each session up to 24 hours after it was initially posted. Participants were asked not to respond unless they had had a full night's sleep between sessions. Email reminders were sent out at pre-set intervals to participants who had not yet completed the current session. Payment for each round was $1.50, with a $1.50 bonus for completing all 4 rounds. The experiment had a retention rate of 70%, with 83 participants completing all 4 rounds. Data from a further 11 participants was excluded due to: a) technical issues (*n* = 2) b) non-native English speakers (*n* = 7) (c) previous exposure to a Semitic language (*n* = 2). Therefore, data from 72 participants was used in the final analysis.

2.3 Procedure

The procedure and methodology used in the current experiment was identical to that in Schembri et al. (2016a) except for presentation order. Participants were informed that they would learn an unknown language, but were not told which aspects of the language to attend to. The experiment consisted of three phases: familiarization, training and review. Both visual and auditory stimuli were presented during each phase. An illustration of the familiarization phase is presented in Figure 1.

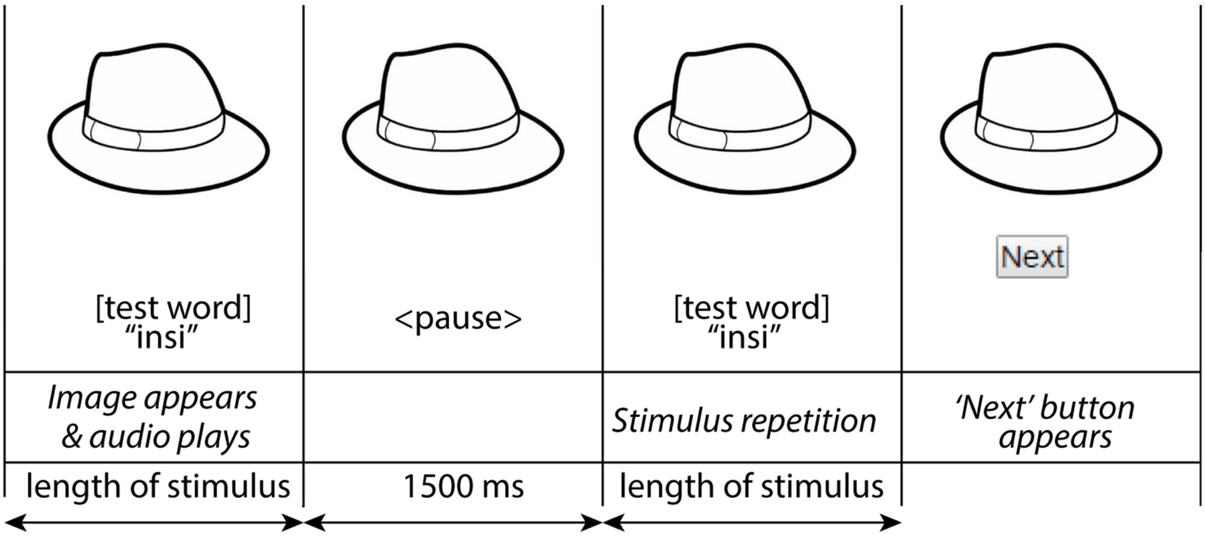


Figure 1: Familiarization phase

Figure 1 demonstrates that, for each trial in the familiarization phase, a word was played back twice, with an image presented simultaneously. Figure 2 below represents the training and generalization phase.

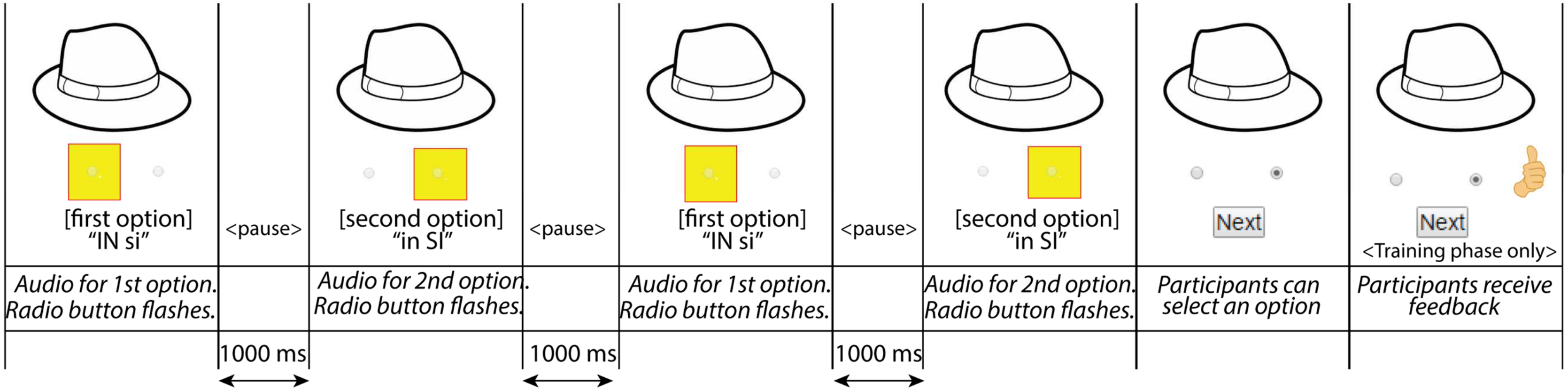


Figure 2: Training and generalization phase. Note that the image remains on screen throughout. It has been removed here for reasons of space. The use of capital letters in this diagram corresponds to an auditorily stressed syllable.

During the training phase, participants were tested by means of a two-alternative forced choice task. Participants were then given feedback on their response. The generalization phase differed from the training phase in two respects: a) participants were tested on novel words which they had not heard previously; b) participants were not given feedback on their response. The first 3 days followed the same sequence, with the exception that days 2 and 3 began with a review of stimuli presented on the previous day(s). This sequence is depicted in Table 1 below. Participants were presented with alternating familiarization and training blocks. After every 2 alternating blocks, a review block was presented, containing stimuli encountered in the preceding blocks. A generalization block completed each day: this contained novel words which had not previously been encountered, but were identical in terms of CV-structure to the words encountered in the familiarization and training phases. The final day (day 4) was not structured into blocks, but simply tested participants' ability to generalize to novel stimuli; these test items represent the full range of variation encountered in the previous 3 days.

Block	Description
Familiarization Block 1	Subjects hear 5 randomly chosen training words.
Training Block 1	Subjects are tested on the 5 words in Familiarization Block 1.
Familiarization Block 2	Subjects hear 5 randomly chosen training words.
Training Block 2	Subjects are tested on the words in Familiarization Block 2.
Review Block 1	Subjects complete a familiarization and training block for all items presented so far (10 items).
Familiarization Block 3	Subjects hear 5 randomly chosen training words.
Training Block 3	Subjects are tested on the words in Familiarization Block 3.
Familiarization Block 4	Subjects hear 5 randomly chosen training words.
Training Block 4	Subjects are tested on the words in Familiarization Block 4.
(Familiarization Block 5)	***This block was present on Day 1 only, in order to replicate the presentation in Schembri et al. (2016a).*** On Day 2 and Day 3, the experiment moved on to Review Block 2.
(Training Block 5)	***This block was present on Day 1 only, in order to replicate the presentation in Schembri et al. (2016a).*** On Day 2 and Day 3, the experiment moved on to Review Block 2.
Review Block 2	Subjects complete a familiarization and training block for all items presented so far. 25 items are reviewed on Day 1. 20 items are reviewed on Day 2 and Day 3.
Generalization Block	Subjects are tested on previously unseen words. Each word in the current day's Familiarization and Training blocks is matched with a novel word which is identical in terms of CV-structure. 25 items are presented on Day 1. 20 items are presented on Day 2 and 3.

Table 1: Experiment structure for days 1-3. Note that days 2 and 3 began with an additional Review Block (not shown), in which words presented on the previous day(s) are reviewed.

Participants were presented with randomly chosen stimuli, with all levels of complexity potentially represented at any point in the experiment. Presentation order was random, with no constraints on the co-occurrence of similar stimuli. Each participant experienced the stimuli in a unique order; because of this, the experiment software stored a list of stimuli presented to each individual participant. This allowed the software to correctly customize the experiment for each participant in a number of ways. First, each stimulus presented in the familiarization and training blocks was paired with a novel stimulus that was identical in terms of CV-structure, to be presented in the generalization block. As each participant was presented with different stimuli in the training and familiarization blocks, the generalization block at the end of each day was similarly unique for each participant. Second, days 2 and 3 began with a review of words seen on previous days. Again, the content of these review blocks was unique for each participant, generated from the stored list of stimuli presented to the participant. Third, the stimuli presented on days 2 and 3 were cross-referenced against the stored list of stimuli for each participant to avoid repeat presentations of stimuli. As a result, while the overall experiment structure was the same for each participant, the presentation order differed in a number of respects. However, by the end of day 3, every participant had been exposed to the same stimuli. Because participants in both experiments were exposed to the same stimuli, and tested on the same set of stimuli on day 4, any differences in performance must be due to the difference in presentation order. If participants in the initially limited input experiment (Schembri et al., 2016a) perform better than those in the randomly ordered experiment, this would support the ‘less is more’ hypothesis in the context of prosodic acquisition. Conversely, if participants in the randomly ordered experiment tested here perform better than those in the previous experiment, this would suggest that the ‘less is more’ hypothesis does not hold for the case of prosodic acquisition.

3. Results and discussion

A generalized linear mixed-effects model was fitted to participant responses in the generalization phase using the lme4 package in R (Bates et al., 2015), which was also used to compute p-values. The model included individual participant responses as the dependent variable, coded as correct (1) or incorrect (0). In order to facilitate comparisons between experiments, we used the same fixed and random factors as in Schembri et al. (2016a). The following fixed effects were included in the model: Constraint Name, and Experiment Day (1-4). Subjects and items were entered as random variables with random intercepts. The formula for the model was Correct ~ Constraint_Name + Day + (1 | Worker_ID) + (1 | Item_Number). Factors were dummy coded, with WSP-CVV as the reference category for the factor Constraint_Name, and day 1 as the reference category for the factor Day. Table 2 presents all main effects with coefficient estimates, standard error, z values and p-values. A supplementary model to check for order effects is included in Appendix A.

	Coefficient	*SE*	*z*	*p*	
(Intercept)	0.889	0.140	6.363	< 0.001	***
EndRule	-0.106	0.158	-0.673	0.501	
EndRule & WSP	0.016	0.159	0.098	0.922	
NonFin	-0.604	0.155	-3.893	< 0.001	***
NonFin & EndRule	-0.155	0.150	-1.031	0.302	
NonFin & EndRule & WSP	0.223	0.157	1.420	0.156	
NonFin & WSP	-0.200	0.150	-1.334	0.182	
Trochaic & A-L	-0.628	0.149	-4.226	< 0.001	***
Trochaic & A-L & EndRule	-0.532	0.152	-3.507	< 0.001	***
Trochaic & A-L & Endrule & Nonfin	-0.428	0.152	-2.809	0.005	**
Trochaic & A-L & NonFin	-1.134	0.145	-7.811	< 0.001	***
WSP	-0.257	0.153	-1.681	0.093	.
WSP3	0.276	0.158	1.746	0.081	.
day2	0.134	0.078	1.710	0.087	.
day3	0.222	0.079	2.820	0.005	**
day4	0.231	0.062	3.737	< 0.001	***

Table 2: Constraints and Experiment Day with coefficient, standard error, z values and p values

The model demonstrates that participants' performance improved steadily throughout the experiment, with better performance on days 3 ($z = 2.820$, $p = 0.005$) and 4 ($z = 3.737$, $p < 0.001$), compared to day 1; a positive effect was also found for day 2, but this was only approaching significance ($z = 1.710$, $p = 0.087$). Participants performed significantly worse than average on words in the following five constraint groups: NONFIN ($z = -3.893$, $p < 0.001$), TROCHAIC & A-L ($z = -4.226$, $p < 0.001$), TROCHAIC & A-L & ENDRULE ($z = -3.507$, $p < 0.001$), TROCHAIC & A-L & ENDRULE & NONFIN ($z = -2.809$, $p = 0.005$), and TROCHAIC & A-L & NONFIN ($z = -7.811$, $p < 0.001$). No significant effects were found for words in the remaining constraint groups. The constraint groups associated with poor performance have an important property in common. Four out of these five constraint groups contain only words in which a light syllable is stressed. The majority of the words in the remaining constraint group, NONFIN, also contain a light stressed syllable. Additionally, there are no words containing light stressed syllables in the remaining constraint groups. Therefore, the results from this model demonstrate that participants were less likely to assign stress to light syllables than heavy or superheavy syllables. In Cairene Arabic, weight drives stress assignment, such that heavy syllables are preferentially stressed over light syllables. However, light syllables are sometimes stressed over heavy syllables due to the complex interaction of multiple constraints. For example, the constraint NONFIN bans stress on the final syllable. Due to the effect of this constraint, a word containing a light and a heavy syllable, such as /'fi.him/ receives stress on a light syllable, even though the word contains a competing heavy syllable. Participants' poor performance on these words indicates that they have acquired the basics of the Cairene Arabic weight system, understanding that light syllables rarely attract stress. However, they have not acquired the exceptional patterns of the language which allow light syllables to be stressed over heavy syllables in specific contexts. Instead, they have acquired an overly general prohibition on stressing light syllables.

The model in Table 2 demonstrates that participants perform poorly on words containing stressed light syllables. However, this is unlikely to be the only factor affecting performance. L1 transfer plays a significant role in L2 acquisition (Anani, 1989; Archibald, 1993; Baptista, 1989;

Mairs, 1989). An analysis of the differences between the English and Cairene Arabic stress system reveals a number of additional factors which may have a significant effect on participants' performance (Schembri et al., 2016a). One of the clearest surface differences between the two stress systems lies in the distribution of stresses in the word. English nouns tend to favour initial stress, and avoid final stress; Cairene Arabic words occur with initial stress less often, and with final stress more often. If participants apply an L1 strategy, they should perform poorly on finally stressed words, and perform well on words with initial stress. Words with final stress are exclusively contained within 2 constraint groups: the constraint group WSPμμμ contains words ending in a stressed superheavy syllable; the constraint group WSP-CVV contains words ending in a CVV syllable. Because of this, it is possible to separate out the effects of these constraint rankings on participants' performance (Schembri et al., 2016a). Additionally, English and Cairene Arabic differ in their treatment of weight. Long vowels and coda consonants contribute equally to stress assignment in Cairene Arabic. However, in English, long vowels are more likely to trigger stress assignment than coda consonants (Guion, et al., 2003; Guion et al., 2004; Guion, 2005). If participants reflect this L1 pattern, the presence of a long vowel should have a significant effect, independent of the effect of weight. In order to test for these interactions, the model included the following factors, taken from Schembri et al. (2016a): Stress Position (initial, medial, final-WSPμμμ, final-WSP-CVV); Weight (Light, Heavy); Long Vowel (Yes, No). The formula for the model was Correct ~ Stress_Position + Weight + Long_Vowel + Day + (1 | Worker_ID) + (1 | Item_Number). The reference category for the factor Stress_Position was Medial; for the factor Weight was Light; for the factor Long_Vowel was long_vowel0 (absence of long vowel). Table 3 presents all main effects with coefficient estimates, standard error, z values and p-values. A supplementary model to check for order effects is included in Appendix A.

	Coefficient	*SE*	*z*	*p*	
(Intercept)	0.402	0.109	3.673	< 0.001	***
stress_posINITIAL	-0.374	0.078	-4.780	< 0.001	***
stress_posFINAL_CVV	-0.124	0.126	-0.985	0.325	
stress_posFINAL_μμμ	0.406	0.156	2.610	0.009	**
long_vowel1	0.358	0.086	4.156	< 0.001	***
weightheavy	0.255	0.083	3.058	0.002	**
day2	0.132	0.078	1.687	0.092	.
day3	0.224	0.079	2.850	0.004	**
day4	0.229	0.062	3.708	< 0.001	***

Table 3: Stress Position, Weight, Long Vowel and Day: coefficient, standard error, z and p values

The model revealed that participants performed best on words with stressed final superheavy syllables (z = 2.610, p = 0.009), and performed worst overall on words which should have had initial stress (z = -4.780, p < 0.001). These results demonstrate that participants are not applying an L1 strategy in terms of stress position, as this would imply the opposite pattern: worse performance on words with stressed final superheavy syllables, and better performance on words with initial stress. In accordance with the results in Table 1, participants were more likely to stress words containing a heavy syllable compared to a light syllable (z = 3.058, p = 0.002). In addition, there was an independent effect for the presence of a long vowel (z = 4.156, p < 0.001). This means that participants were more likely to stress CVV syllables over CVC syllables, although both are classified as heavy; and more likely to stress CVVC syllables over CVCC syllables, although both are classified as superheavy. This result demonstrates an L1 transfer effect in terms of weight. If participants had fully acquired the Cairene Arabic weight system, there should be a positive significant effect for weight, but no independent significant effect for the presence of long vowels. In other words, participants should be equally likely to stress CVV and CVC syllables (both heavy); and CVVC and CVCC syllables (both superheavy). The model also demonstrated that performance improved throughout the experiment, confirming the results in Table 1: participants performed significantly better on days 3 ($z = 2.850, p = 0.004$) and 4 ($z = 3.708, p = < 0.001$); additionally, a positive effect approaching significance was found for day 2 ($z = 1.687, p = 0.092$).

The overall trajectory of participants' performance in the current experiment is strikingly different to that in the staged input experiment (Schembri et al., 2016a). In the current experiment, participants' performance continued to improve throughout. In the staged input experiment, participants' performance was best on the first day of the experiment, when adjusted for significant factors affecting performance. Figure 3 presents correct responses for the generalization phase on each day of both the staged input (Schembri et al., 2016a) vs. random input experiments, averaged across participants.

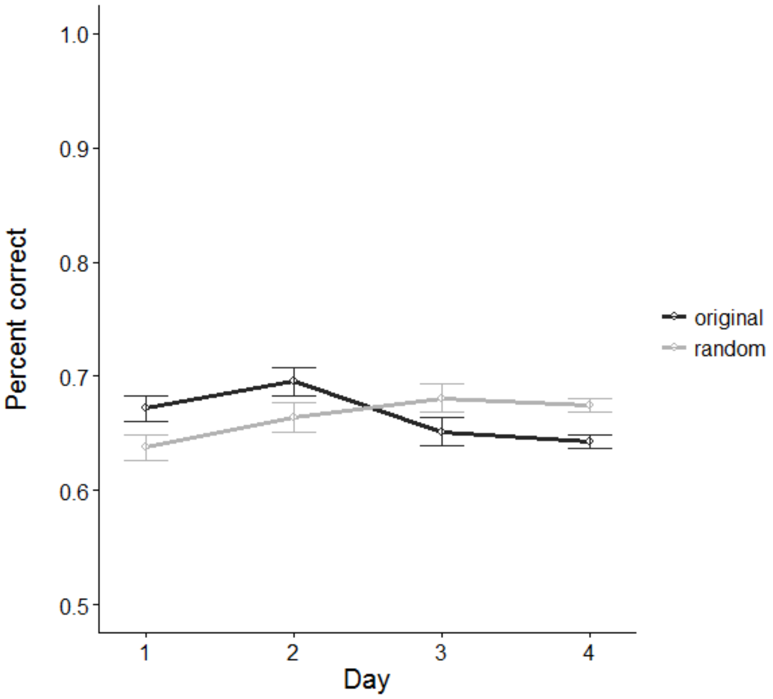


Figure 3: Correct responses for the generalization phase on each day, averaged over participants. Error bars represent standard error.

Thus, participants in the staged input experiment initially performed better than those in the random experiment; however, performance dropped midway through the experiment. It is particularly striking that participants in the random experiment begin at a lower level of performance than those in the staged input experiment, but then surpass them, perhaps more closely simulating a ‘real world’ learning situations, rather than a staged ‘classroom’ experience.

Participants’ patterns of acquisition are also quite different across the two experiments. In order to allow for a direct comparison between experiments, data from day 4 across both experiments was combined into a single dataset. The final day of the experiment is ideal for comparing participants’ performance. Day 4 was devoted exclusively to generalization and included novel stimuli for all word types introduced on the previous 3 days. Performance on this day is the most significant indicator of overall performance. Furthermore, participants in both experiments were exposed to the same set of stimuli on this day, facilitating a direct comparison of performance. As a result, any differences in performance can only be attributed to the differences in presentation order on days 1-3. An additional model was run on this dataset, adding an interaction between experiment type (original or random) and each of the factors in the model. The formula for the model was Correct ~ Experiment * Long_Vowel + Experiment * Stress_Pos + Experiment * Weight + (1 | Worker_ID) + (1 | Item_Number). The reference category for the factor Experiment was Original. Table 4 presents all main effects with coefficient estimates, standard error, z values and p-values. A similar dataset was created for the OT-constraint model; however, this produced inconclusive results. This analysis is included in Appendix A. An additional dataset was created for the surface stress model for all 4 days, showing similar results and corroborating the conclusions drawn from Figure 3. This is included in Appendix A.

	Coefficient	*SE*	*z*	*p*	
experiment_random	-0.110	0.133	-0.830	0.407	
long_vowel1	0.106	0.103	1.024	0.306	
stress_posINITIAL	-0.366	0.097	-3.780	< 0.001	***
stress_posFINAL_CVV	-0.867	0.151	-5.743	< 0.001	***
stress_posFINAL_μμμ	-0.465	0.196	-2.366	0.018	*
weightheavy	0.159	0.107	1.485	0.138	
experiment_random:long_vowel1	0.121	0.116	1.047	0.295	
experiment_random:stress_posINITIAL	-0.001	0.106	-0.014	0.989	
experiment_random: stress_posFINAL_CVV	0.761	0.171	4.441	< 0.001	***
experiment_random: stress_posFINAL_μμμ	0.979	0.221	4.424	< 0.001	***
experiment_random:weightheavy	0.305	0.117	2.611	0.009	**

Table 4: Combined dataset (day 4 only): coefficients, standard error, z values and p values

The conclusion that participants in the random experiment acquired the Cairene Arabic stress system more successfully than those in the staged input experiment is strengthened by examining the additional differences between participants in the two experiments. Participants in the random experiment outperformed those in the staged input experiment on words with final superheavy syllables ($z = 4.441$, $p < 0.001$), and words ending in a CVV syllable ($z = 4.424$, $p < 0.001$). In addition, participants in the random experiment were more likely to stress heavy rather than light syllables ($z = 2.611$, $p = 0.009$). Participants' performance on words with final stress diverged significantly across the two experiments. Participants in the current experiment performed significantly better than those in the staged input experiment on both sets of words with final stress: those in the constraint group WSPμμμ, as well as those in the constraint group WSP-CVV. However, this effect is most obvious in words in the constraint group WSPμμμ. These words were associated with best performance overall for participants in the current experiment. In contrast participants in the staged input experiment performed best overall on these words in training, but not in generalization. Figure 4 shows participants' performance on words with each main stress position in training and generalization in the staged input experiment. Figure 5 shows the corresponding graph for the random experiment. In these figures, the category 'medial stress' includes words with non-final stress on both the second and third syllable, as there was no difference in performance between these two groups.

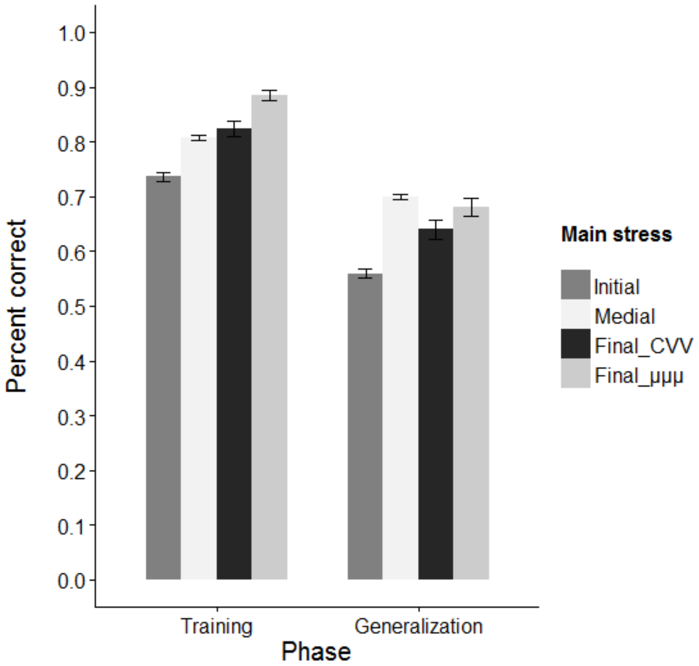


Figure 4: Staged input experiment: performance by main stress position. Error bars represent standard error.

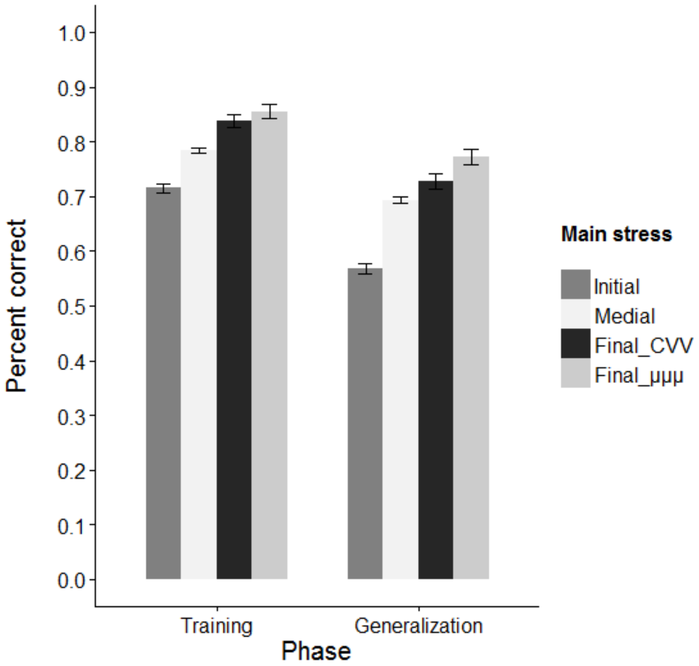


Figure 5: Random experiment: performance by main stress position. Error bars represent standard error.

Overall trends should remain the same in training and generalization. For example, in both experiments, participants perform better on words with long vowels than words with no long vowels, in both the training and generalization phase. However, Figure 4, which shows participants' performance by main stress position in the staged input experiment, illustrates a case in which the direction of an effect is reversed in generalization: in the training phase, participants perform better on words with final stress compared to medial stress; in the generalization phase, participants perform better on words with medial stress compared to final stress. Figure 5 demonstrates that this striking pattern of reversal seen in words with medial and final stress in the staged input experiment is no longer present in the random experiment. Words with final stress are acquired in training in both experiments. However, in the current experiment, they are retained and mirrored more closely in the generalization phase. It is unclear why the changes in presentation order across the two experiments have had this effect.

Participants across the two experiments also differed in their acquisition of the underlying Cairene Arabic weight system. In both experiments, the presence of a long vowel had a significant effect on stress assignment. Participants were significantly more likely to stress a syllable containing a long vowel. However, weight (light vs heavy syllables) did not affect the stress assignments strategies of participants in the staged input experiment, independent of the effect of long vowels. That is, while participants were significantly more likely to stress any syllable containing a long vowel, they were equally likely to stress CV and CVC syllables. In other words, the presence of a coda consonant had no effect on the stress assignment strategies of participants in the staged input experiment. Conversely, the presence of coda consonants had a strong significant effect on the stress assignment strategies of participants in the current experiment, who were significantly more likely to stress heavy than light syllables. Nevertheless, the presence of a long vowel had a significant effect on the stress assignment strategies of both sets of participants, indicating that neither set of participants fully acquired the underlying Cairene Arabic weight system. If the underlying weight system had been fully acquired, the presence of a long vowel should not have an effect independent of the effect of weight. That is, CVV and CVC syllables, and CVVC and CVCC syllables, should have an equal effect on stress assignment.

Overall, participants in the current experiment demonstrated a lessened L1 transfer effect compared to participants in the staged input experiment, both in terms of the distribution of stresses in the word, as well as in their acquisition of the underlying Cairene Arabic weight system: they performed better on words with final stress, which is uncommon in English nouns. Additionally, they acquired a better understanding of the weight system, such that they were more likely to stress CVC syllables over CV syllables. Taken together, these results indicate that participants in the current randomized order of presentation experiment had a more accurate understanding of the underlying prosodic structure and organization of the language.

4. Conclusion

According to the “less is more” hypothesis, children’s limited working memory capacity is helpful in learning language; therefore, adults’ relative difficulty in acquiring language in later life may be due, in part, to their increased cognitive capabilities. This hypothesis has been tested primarily by directly comparing adults’ acquisition when input is initially limited, compared to when they are initially exposed to the full range of complexity of the target language. Studies which have demonstrated an advantage for initially limited input, whether in terms of stimulus length or grammatical complexity, have been taken as evidence for the ‘less is more’ hypothesis. The current study is the first comparison of this kind in the domain of prosodic acquisition. The results demonstrate that participants in the random experiment outperformed participants in the staged input experiment overall, as well as on all metrics measuring their knowledge of aspects of the underlying prosodic structure of the language. As a result, we conclude that, at least for the narrow domain of main word stress acquisition, the ‘less is more’ hypothesis does not hold.

These results may be narrowly applicable to prosodic acquisition due to its unique characteristics. Shorter words may not contain enough information to correctly induce the stress (or tonal) pattern of a language, particularly if the system is complex. For example, given input containing only bisyllabic words with initial stress, it is impossible to determine which of a number of competing hypotheses is correct. These possible hypotheses include, but are not limited to: invariant initial stress, invariant penultimate stress, predictable stress based on quantity sensitivity, lexical stress, and so on. The learner can begin to rule out hypotheses only once they are presented with a variety of longer words. For example, once the input contains three syllable words, it becomes possible to rule out either invariant initial stress, or invariant penultimate stress. Therefore, artificially limiting the initial complexity of the input may harm participants’ ability to form correct hypotheses about the stress system. This possibility could be tested by replicating the current experiment with the stress systems of other languages. The learning of grammatical tone systems, as found in many Bantu languages (Hyman & Kisseberth, 1998) may also be impossible to deduce without reference to longer words (cf. Demuth, 1993). If participants acquiring a variety of stress and tone systems demonstrate an advantage for input which is initially varied in terms of complexity, while participants acquiring other systems demonstrate the opposite advantage, this would be evidence that the ‘less is more’ hypothesis does not apply equally to the acquisition of all linguistic domains.

Appendix A: Supplementary Statistics

Descriptive Statistics

Each of the factors included in the statistical models in Tables 2 and 3 are illustrated through bar graphs, plotted against accuracy, below.

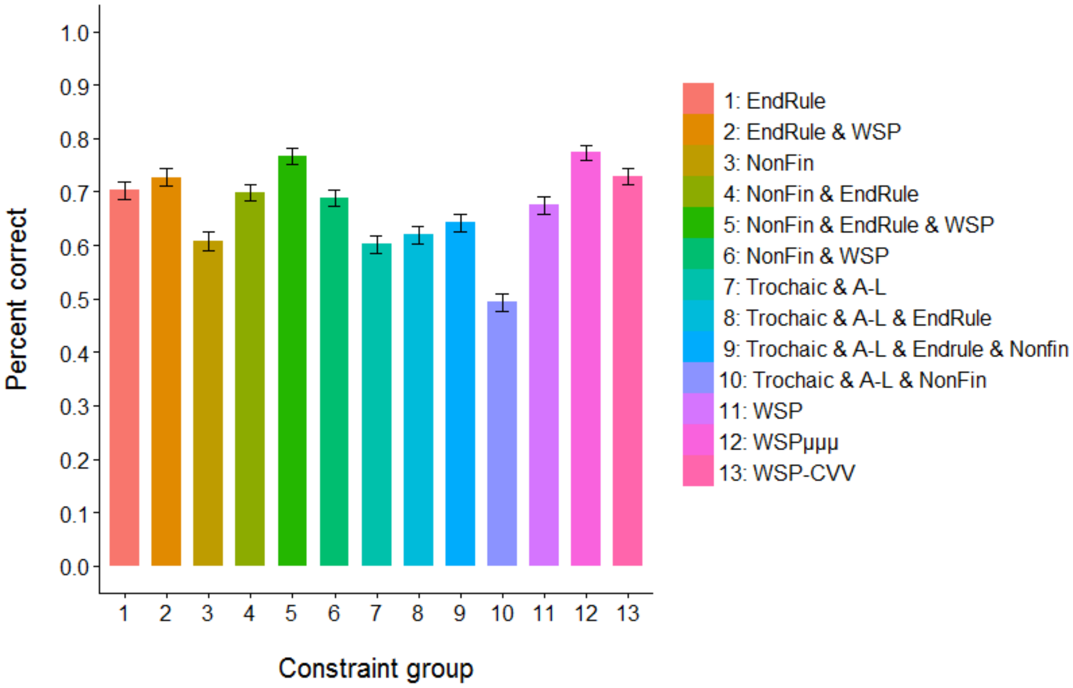


Figure A1: Correct responses by constraint group. Error bars represent standard error.

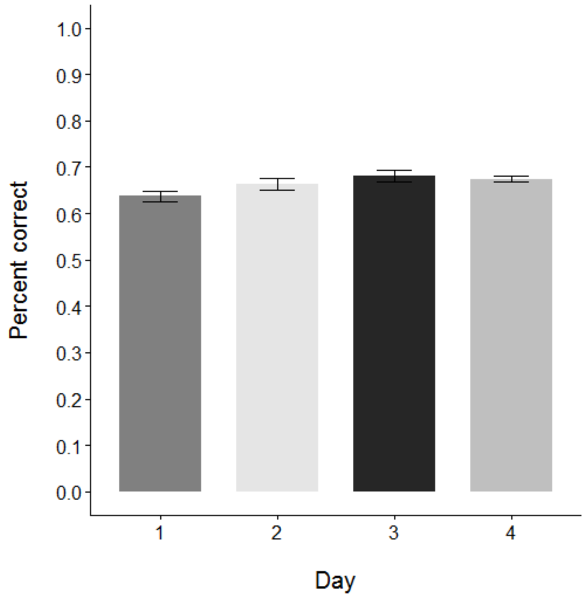


Figure A2: Correct responses by day. Error bars represent standard error.

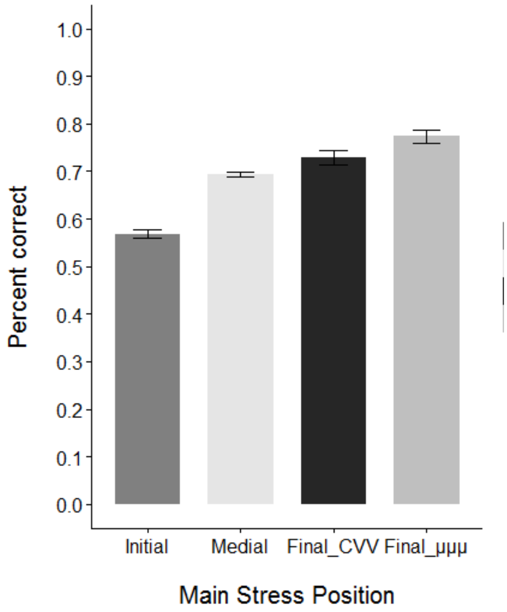


Figure A3: Correct responses by main stress position. Error bars represent standard error.

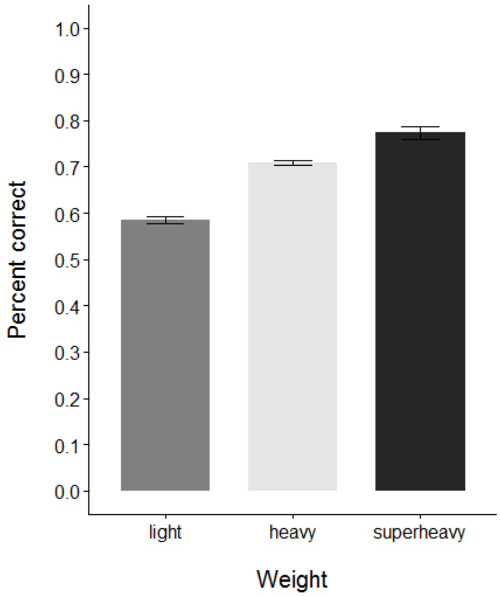


Figure A4: Correct responses by weight. Error bars represent standard error.

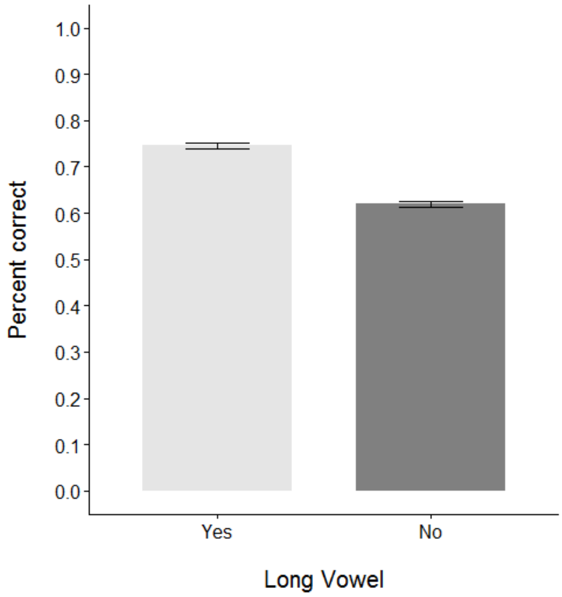


Figure A5: Correct responses by presence of long vowel. Error bars represent standard error.

Order Effects

In order to check for order effects, an additional factor ‘Correct_Syll_Order’ was added to the statistical models in Tables 2 and 3. This factor encodes whether the correct choice was presented as the first or second item. Table A1 replicates the statistical model in Table 2 with the inclusion of ‘Correct_Syll_Order’. The formula for this model is Correct ~ Constraint_Name + Day + Correct_Syll_Order + (1 | Worker_ID) + (1 | Item_Number). Table A2 replicates the statistical model in Table 3 with the inclusion of ‘Correct_Syll_Order’. The formula for this model is Correct ~ Stress_Position + Weight + Long_Vowel + Day + Correct_Syll_Order + (1 | Worker_ID) + (1 | Item_Number).

	Coefficient	*SE*	*z*	*p*	
(Intercept)	0.920	0.155	5.920	0.000	***
EndRule	0.360	0.189	1.904	0.057	.
EndRule & WSP	0.632	0.198	3.196	0.001	**
NonFin	-0.438	0.186	-2.348	0.019	*
NonFin & EndRule	0.536	0.185	2.897	0.004	**
NonFin & EndRule & WSP	0.615	0.192	3.208	0.001	**
NonFin & WSP	0.374	0.187	1.998	0.046	*
Trochaic & A-L	-0.167	0.184	-0.906	0.365	
Trochaic & A-L & EndRule	0.056	0.189	0.297	0.767	
Trochaic & A-L & Endrule & Nonfin	0.202	0.191	1.055	0.291	
Trochaic & A-L & NonFin	-0.318	0.178	-1.782	0.075	.
WSP	0.032	0.182	0.177	0.859	
WSPμμμ	0.188	0.186	1.009	0.313	
day2	-0.248	0.125	-1.984	0.047	*
day3	-0.237	0.123	-1.928	0.054	.
day4	-0.270	0.085	-3.186	0.001	**
correct_syll_order2	-0.279	0.042	-6.652	0.000	***

Table A1: Constraints, Experiment Day and Correct Syllable Order with coefficient, standard error, z values and p values

	Coefficient	*SE*	*z*	*p*	
(Intercept)	0.430	0.112	3.847	0.000	***
stress_posINITIAL	-0.374	0.078	-4.787	0.000	***
stress_posFINAL_CVV	-0.125	0.126	-0.991	0.322	
stress_posFINAL_μμμ	0.405	0.155	2.607	0.009	**
long_vowel1	0.358	0.086	4.169	0.000	***
weightheavy	0.253	0.083	3.043	0.002	**
day2	0.130	0.078	1.658	0.097	.
day3	0.222	0.079	2.824	0.005	**
day4	0.227	0.062	3.685	0.000	***
correct_syll_order2	-0.052	0.042	-1.236	0.216	

Table A2: Stress Position, Weight, Long Vowel, Day and Correct Syllable Order: coefficient, standard error, z and p values

The statistical models in Tables A1 and A2 demonstrate that there were no significant order effects. This indicates that participants did not have any bias towards selecting either the first option or the second option.

Word Length

In order to check whether word length had a significant effect on participants' performance, an additional factor 'Num_Sylls' was added to the statistical models in Tables 3 and 4. This factor is a continuous variable which encodes how long each word is, in terms of number of syllables. Table

A3 replicates the statistical model in Table 2 with the inclusion of ‘Num_Sylls’. The formula for this model is Correct ~ Constraint_Name + Day + Num_Sylls + (1 | Worker_ID) + (1 | Item_Number). Table A4 replicates the statistical model in Table 3 with the inclusion of ‘Num_Sylls’. The formula for this model is Correct ~ Stress_Position + Weight + Long_Vowel + Day + Num_Sylls + (1 | Worker_ID) + (1 | Item_Number).

	Coefficient	*SE*	*z*	*p*	
(Intercept)	0.611	0.278	2.201	0.028	*
EndRule	0.329	0.193	1.700	0.089	.
EndRule & WSP	0.535	0.235	2.274	0.023	*
NonFin	-0.392	0.195	-2.010	0.044	*
NonFin & EndRule	0.506	0.194	2.604	0.009	**
NonFin & EndRule & WSP	0.529	0.223	2.373	0.018	*
NonFin & WSP	0.352	0.189	1.863	0.063	.
Trochaic & AFL	-0.145	0.185	-0.780	0.435	
Trochaic & AFL & EndRule	-0.026	0.221	-0.119	0.905	
Trochaic & AFL & Endrule & Nonfin	0.120	0.223	0.537	0.591	
Trochaic & AFL & NonFin	-0.337	0.180	-1.871	0.061	.
WSP	0.015	0.183	0.082	0.935	
WSP3	0.178	0.188	0.944	0.345	
day2	-0.252	0.125	-2.018	0.044	*
day3	-0.240	0.123	-1.952	0.051	.
day4	-0.273	0.085	-3.212	0.001	**
num_sylls	0.062	0.088	0.706	0.480	

Table A3: Constraints, Experiment Day and Number of Syllables with coefficient, standard error, z values and p values

	Coefficient	*SE*	*z*	*p*	
(Intercept)	0.414	0.265	1.565	0.117	
stress_posFINAL_μμμ	0.184	0.194	0.952	0.341	
stress_posINITIAL	-0.040	0.184	-0.219	0.827	
stress_posMEDIAL	0.458	0.170	2.693	0.007	**
long_vowel1	0.210	0.108	1.947	0.051	.
weightlight	-0.088	0.119	-0.736	0.462	
day2	-0.122	0.115	-1.060	0.289	
day3	-0.214	0.120	-1.790	0.073	.
day4	-0.225	0.083	-2.725	0.006	**
num_sylls	0.048	0.076	0.625	0.532	

Table A4: Stress Position, Weight, Long Vowel, Day and Number of Syllables: coefficient, standard error, z and p values

The above models demonstrate that word length is not a significant factor affecting participants' performance. This holds true both if participants are acquiring OT constraint rankings, as well as if participants are acquiring surface level stress patterns.

Combined dataset: OT-constraint model

In order to allow for a direct comparison between experiments, data from day 4 across both experiments was combined into a single dataset. Participants in both experiments were exposed to the same set of stimuli on this day, facilitating a direct comparison of performance. This was completed for the surface stress model in Table 4. A similar comparison was attempted for the OT constraint model. However, this model did not converge, indicating that there is insufficient data to fit a model to the data. This is because the OT-constraint model is significantly more complex than the surface stress model, with many more degrees of freedom. Therefore, data from a much larger pool of participants is required in order to draw meaningful conclusions. Despite these limitations, the model was forced to converge by doubling the number of iterations allowed to fit a model. It was then possible to fit a model to the data. However, this strategy may cause the model to draw incorrect conclusions. Therefore, the results from this model should be interpreted with caution. An interaction between experiment type (original or random) and each of the factors in the OT-constraint model was included. The formula for the model was Correct ~ Experiment * Constraint_Name + (1 | Worker_ID) + (1 | Item_Number). The reference category for the factor Experiment was Original. Table A5 presents all main effects with coefficient estimates, standard error, z values and p-values.

	Coefficient	*SE*	*z*	*p*	
(Intercept)	0.096	0.140	0.687	0.492	
experiment_random	1.073	0.177	6.051	< 0.001	***
EndRule	0.710	0.177	4.011	< 0.001	***
EndRule & WSP	1.106	0.182	6.067	< 0.001	***
NonFin	0.525	0.175	2.994	0.003	**
NonFin & EndRule	0.898	0.163	5.517	< 0.001	***
NonFin & EndRule & WSP	1.055	0.174	6.059	< 0.001	***
NonFin & WSP	0.598	0.169	3.532	< 0.001	***
Trochaic & AFL	0.357	0.163	2.186	0.029	*
Trochaic & AFL & EndRule	0.506	0.169	3.001	0.003	**
Trochaic & AFL & Endrule & Nonfin	0.628	0.177	3.550	< 0.001	***
Trochaic & AFL & NonFin	0.028	0.155	0.183	0.854	
WSP	0.551	0.169	3.261	0.001	**
WSPμμμ	0.256	0.182	1.405	0.160	
experiment_random:EndRule	-0.780	0.216	-3.606	< 0.001	***
experiment_random:EndRule & WSP	-1.162	0.221	-5.266	< 0.001	***
experiment_random:NonFin	-1.251	0.208	-6.008	< 0.001	***
experiment_random:NonFin & EndRule	-1.097	0.198	-5.551	< 0.001	***
experiment_random: NonFin & EndRule & WSP	-0.704	0.215	-3.270	0.001	**
experiment_random:NonFin & WSP	-0.925	0.201	-4.592	< 0.001	***
experiment_random:Trochaic & AFL	-1.045	0.196	-5.336	< 0.001	***

experiment_random: Trochaic & AFL & EndRule	-1.156	0.201	-5.737	< 0.001	***
experiment_random: Trochaic & AFL & Endrule & Nonfin	-1.170	0.208	-5.620	< 0.001	***
experiment_random: Trochaic & AFL & NonFin	-1.213	0.187	-6.493	< 0.001	***
experiment_random:WSP	-0.863	0.204	-4.238	< 0.001	***
experiment_random:WSPμμμ	0.094	0.222	0.423	0.672	

Table A5: Combined dataset (OT-constraint model): coefficients, standard error, z values and p values

The results shown in this model are difficult to interpret. According to the model, participants performed significantly worse in the random experiment, as compared to the original experiment, for all constraint groups other than WSPμμμ. However, these results can be shown to be false by referring to the raw data. For example, the model states that participants in the original experiment outperformed those in the random experiment on words in the constraint group EndRule ($z = -0.780$, $p < 0.001$). However, participants in the original experiment scored 68% correct on these words, while participants in the random experiment scored 73%. The discrepancies between the results in Table A5 and the raw data are further illustrated in Table A6 below.

	Original	*Random*
CVV#	52%	**74%**
EndRule	68%	**73%**
EndRule & WSP	**76%**	73%
NonFin	**64%**	60%
NonFin & EndRule	**72%**	71%
NonFin & EndRule & WSP	74%	**80%**
NonFin & WSP	66%	**68%**
Trochaic & AFL	60%	60%
Trochaic & AFL & EndRule	**63%**	61%
Trochaic & AFL & Endrule & Nonfin	**67%**	64%
Trochaic & AFL & NonFin	**53%**	50%
WSP	65%	**68%**
WSPμμμ	58%	**80%**

Table A6: Percent correct for each constraint group in both original and random experiment. Higher scores are marked in bold.

Table A6 demonstrates that participants perform better on words in 6 constraint groups in both the original and random experiment, with one constraint group (Trochaic & AFL) demonstrating identical performance in both experiments. As these raw figures diverge so dramatically from those in the statistical model in Table A5, and given that this model was unable to converge under normal parameters, it seems reasonable to conclude that there is insufficient data to accurately fit the model in Table A5 to the data.

Combined dataset: Surface stress model (all 4 days)

Data from all 4 days, across both experiments, was combined into a single dataset. This analysis complements the analysis included in Table 4. However, this analysis does not allow for a direct comparison of performance across the same set of stimuli, as participants in the two experiments were exposed to different stimuli on days 1-3, which may affect performance in unpredictable ways. Please refer to Table 4 for a direct comparison of participants' performance on the same set of stimuli. An interaction between experiment type (original or random) and each of the factors in the surface stress model in Table 3, as well as the factor Day, was included. The formula for the model was Correct ~ Experiment * Long_Vowel + Experiment * Stress_Pos + Experiment * Weight + (1 | Worker_ID) + (1 | Item_Number). The reference category for the factor Experiment was Original. Table A7 presents all main effects with coefficient estimates, standard error, z values and p-values.

	Coefficient	*SE*	*z*	*p*	
(Intercept)	0.910	0.126	7.199	< 0.001	***
experiment_random	-0.519	0.150	-3.462	0.001	***
long_vowel1	0.235	0.089	2.638	0.008	**
stress_posINITIAL	-0.565	0.084	-6.731	< 0.001	***
stress_posFINAL_CVV	-0.476	0.130	-3.652	< 0.001	***
stress_posFINAL_μμμ	-0.241	0.168	-1.431	0.152	
weightheavy	0.060	0.093	0.647	0.517	
day2	-0.039	0.095	-0.409	0.682	
day3	-0.158	0.101	-1.561	0.119	
day4	-0.166	0.071	-2.334	0.020	*
experiment_random: long_vowel1	0.105	0.087	1.202	0.229	
experiment_random: stress_posINITIAL	0.212	0.080	2.659	0.008	**
experiment_random: stress_posFINAL_CVV	0.334	0.130	2.567	0.010	*
experiment_random: stress_posFINAL_μμμ	0.641	0.169	3.786	< 0.001	***
experiment_random: weightheavy	0.206	0.090	2.298	0.022	*
experiment_random: day2	0.170	0.123	1.381	0.167	

experiment_random: day3	0.386	0.129	3.005	0.003	**
experiment_random: day4	0.403	0.090	4.471	< 0.001	***

Table A7: Combined dataset (all 4 days): coefficients, standard error, z values and p values

When considering all 4 days of the experiment, participants' performance was significantly worse in the random experiment, as compared to the staged input experiment ($z = -0.519$, $p = 0.001$). This can be explained through Figure 3, which shows that participants in the random experiment initially performed significantly worse than those in the staged input experiment, overtaking them in performance at a later stage. Importantly, Table 4, which compares performance on day 4, does not show a significant different in overall performance across the two experiments. Because participants on days 1-3 were exposed to different sets of stimuli in the two experiments, direct comparison on participants' performance across the two experiments in necessarily noisier and less reliable than the analysis found in Table 4. Nevertheless, these results broadly support the conclusions drawn from the analysis in Table 4, similarly finding that participants in the random experiment outperformed those in the staged input experiment on words with final superheavy syllables ($z = 0.641$, $p < 0.001$), words ending in a CVV syllable ($z = 0.334$, $p = 0.01$), and words with a heavy syllable ($z = 0.206$, $p = 0.022$). Additionally, this analysis also finds that participants in the random experiment outperformed those in the staged input experiment on words with initial stress ($z = 0.212$, $p = 0.008$). This would suggest that participants in the random experiment were less influenced by L1 transfer. However, as the analysis in Table 4 finds no significant effect for words with initial stress, this result should be interpreted with caution.

The most important finding from this analysis is that participants in the random experiment significantly outperformed those in the staged input experiment on days 3 ($z = 0.386$, $p = 0.003$) and 4 ($z = 0.403$, $p < 0.001$), with the largest effect size found on day 4. This corroborates the results demonstrated in Figure 3, which shows that participants in the staged input experiment initially performed better than those in the random experiment, with performance dropping on day 3. Participants in the random experiment began at a lower level of performance, but steadily improved throughout the experiment, surpassing participants in the staged input experiment on day 3. The results in Table A7 reflect those seen in Figure 3, suggesting that this interpretation is correct.

Appendix B: Stimuli

Constraint group WSP

Familiarization Phase

'taa.wa	HL
'in.si	HL
ga.'lam.bu	LHL
di.'raa.si	LHL
wa.gi.'bat.li	LLHL

Generalization Phase

'faa.tu	HL
'ʃat.mu	HL
ma.'gaa.lu	LHL
bu.'lan.di	LHL
di.na.'mii.ki	LLHL

Generalization Phase (Day 4)

'ruk.ni	HL
'ʃuu.fi	HL
a.'saa.mi	LHL
ka.'lam.na	LHL
ku.tu.'muu.tu	LLHL

Constraint group WSPµµµ

__Familiarization Phase__

mus.tan.za.'maat	HHLS
it.fat.'wint	HHS
baʃ.'niin	HS
ma.ba.'namʃ	LLS
za.la.'laan	LLS

__Generalization Phase__

ma.tin.ʃi.'gilʃ	LHLS
mab.tin.'zilʃ	HHS
sa.'fruut	HS
wa.fa.'daan	LLS
sa.ka.'lans	LLS

__Generalization Phase (Day 4)__

taʃ.ta.ra.'waan	HHLS
it.nam.'fizt	HHS
tan.'ziim	HS
ta.ra.'biiz	LLS
fu.ru.'sint	LLS

Constraint group NONFIN

Familiarization Phase

'rik.bit	HH
'maa.ziz	HH
'ba.nit	LH
'za.mat	LH
'ki.niz	LH

Generalization Phase

'ʃuf.lak	HH
'ʃaa.win	HH
'ma.san	LH
'wa.jam	LH
'ga.bad	LH

Generalization Phase (Day 4)

'bad.ris	HH
'gaa.lis	HH
'fi.rig	LH
'ʃa.naf	LH
'da.mak	LH

Constraint group ENDRULE-R

Familiarization Phase

is.'tab.da	HHL
as.'wan.li	HHL
ʃam.'bat.li	HHL
gib.'tuu.li	HHL
da.waʃ.'naa.ku	LHHL

Generalization Phase

um.'baa.ʃi	HHL
ig.'maa.li	HHL
bar.'kit.li	HHL
gin.'sii.tu	HHL
mi.nas.'baa.lu	LHHL

Generalization Phase (Day 4)

it.'naa.da	HHL
ab.'rii.mi	HHL
mis.'tas.ni	HHL
tag.'rii.di	HHL
za.man.'kaa.wi	LHHL

Constraint group WSP-CVV

Familiarization Phase

fak.'raa	HH
lab.'saa	HH
ʃuf.'tii	HH
bi.tin.'saa	LHH
ba.la.'dii	LLH

Generalization Phase

taf.'taa	HH
jiʃ.'fii	HH
ʃin.'waa	HH
ʃa.tam.'tii	LHH
da.fa.'nuu	LLH

Generalization Phase (Day 4)

gib.'naa	HH
nas.'jaa	HH
gam.'bii	HH
fa.raʃ.'naa	LHH
mu.ba.'laa	LLH

Constraint group ENDRULE-R & WSP

Familiarization Phase		*Generalization Phase*	
mis.til.'mii.nu	HHHL	ban.tis.'ban.ja	HHHL
an.ti.'kan.ja	HLHL	jus.ta.'fan.di	HLHL
iʃ.ti.'rii.li	HLHL	tis.ta.'fii.du	HLHL
kis.ti.'naa.wi	HLHL	muk.li.'maa.ni	HLHL
bi.jif.tik.'ruu.ni	LHLHL	mi.bas.ba.'saa.ti	LHLHL

Generalization Phase (Day 4)

is.tif.'zaa.zi	HHHL
in.ti.'waa.zi	HLHL
ik.ti.'bii.li	HLHL
taʃ.ri.'faa.ti	HLHL
bi.jiʃ.ti.'kii.li	LHLHL

Constraint group TROCHAIC & ALIGN-L

Familiarization Phase / ***Generalization Phase***

Familiarization Phase		Generalization Phase	
'za.ki	LL	'gi.za	LL
'bi.la	LL	'ka.ba	LL
'ga.tu	LL	'sa.da	LL
'lu.ta.ri	LLL	'di.na.mu	LLL
'za.ga.li	LLL	'nu.ka.ti	LLL

Generalization Phase (Day 4)

'du.ga	LL
'ʃi.fa	LL
'dʒi.li	LL
'ga.ma.li	LLL
'sa.ka.ni	LLL

Constraint group NONFIN & WSP

Familiarization Phase

ti.'nas.bak	LHH
mi.'kam.fit	LHH
mi.'zam.bin	LHH
da.'faa.tir	LHH
li.'saa.nik	LHH

Generalization Phase

bi.'nis.nid	LHH
di.'ras.tak	LHH
la.'zaz.tik	LHH
la.'daa.jin	LHH
wi.'laa.dak	LHH

Generalization Phase (Day 4)

mi.'tam.fis	LHH
mi.'faz.lik	LHH
ta.'san.sun	LHH
mi.'gaa.nis	LHH
fi.'luu.sak	LHH

Constraint group NONFIN & ENDRULE-R

Familiarization Phase

an.'faa.sik	HHH
saj.'bin.lik	HHH
is.'tad.sim	HHH
gi.lig.'naa.jit	LHHH
sa.la.'mit.kum	LLHH

Generalization Phase

mat.'saa.fin	HHH
it.'ʃan.kam	HHH
mit.'bas.tif	HHH
bi.tit.'naa.win	LHHH
ta.ta.'naa.sab	LLHH

Generalization Phase (Day 4)

it.'kas.kis	HHH
bit.'zaa.kin	HHH
mis.'taʃ.kil	HHH
ta.man.'taa.ʃan	LHHH
mu.ta.'faa.wit	LLHH

Constraint group TROCHAIC & ALIGN-L & ENDRULE-R & PARSE-FT

Familiarization Phase		*Generalization Phase*	
mit.fat.'fi.ta	HHLL	hat.ban.'li.na	HHLL
aw.fan.'li.na	HHLL	mit.ban.'ʃi.ma	HHLL
a.gib.'lu.ku	LHLL	sa.ban.'si.gi	LHLL
mu.zak.'ri.tu	LHLL	ba.rik.'ti.lu	LHLL
bi.tiʃ.'ra.bu	LHLL	mi.dam.'bi.ka	LHLL

Generalization Phase (Day 4)

is.tag.'fi.ru	HHLL
mus.tam.'ti.ka	HHLL
ba.naf.'si.gi	LHLL
bi.tim.'si.ki	LHLL
ti.ʃuf.'li.na	LHLL

Constraint group TROCHAIC & ALIGN-L & NONFIN

Familiarization Phase		*Generalization Phase*	
'ka.ti.fan	LLH	'ta.ra.lan	LLH
'ʃa.ra.dit	LLH	'a.ba.dan	LLH
'na.za.fit	LLH	'ki.ra.win	LLH
'ma.za.lan	LLH	'na.ʃa.zit	LLH
'ma.ba.sak	LLH	'ba.la.dak	LLH

Generalization Phase (Day 4)

'ka.ta.bit	LLH
'ka.ʃa.fit	LLH
'fa.ʃa.lit	LLH
'da.ra.sit	LLH
'ra.ka.sit	LLH

Constraint group NONFIN & ENDRULE-R & WSP

Familiarization Phase

maʃ.ta.'rak.tiʃ	HLHH
iʃ.ta.'gan.tum	HLHH
tim.bi.'sii.lik	HLHH
mis.ti.'baa.rik	HLHH
in.ti.'zaa.rik	HLHH

Generalization Phase

is.ta.'ʃan.tak	HLHH
is.bi.'laa.jit	HLHH
mus.ta.'waa.kum	HLHH
jin.gu.'duu.lak	HLHH
il.ma.'baa.lig	HLHH

Generalization Phase (Day 4)

is.ti.'kan.tik	HLHH
if.ta.'kan.tak	HLHH
baʃ.ti.'kii.lak	HLHH
mis.ti'kaa.min	HLHH
ib.ti.'daa.kan	HLHH

Constraint group ALIGN-L & ENDRULE-R & NONFIN

Familiarization Phase

ha.niʃ.'ta.gal	LHLH
ka.tab.'ti.lik	LHLH
bi.nak.'ta.sin	LHLH
ma.ʃuf.'tu.kum	LHLH
mus.tak.'ba.lan	LHLH

Generalization Phase

bi.tit.'ki.sif	LHLH
ha.nig.'si.bik	LHLH
ji.gib.'lu.kum	LHLH
bi.tak.'ta.mid	LHLH
ʃa.tam.'tu.kum	LHLH

Generalization Phase (Day 4)

mu.kal.'mi.tak	LHLH
bi.tit.'li.bis	LHLH
ha.tit.'ki.ʃif	LHLH
bi.jan.'da.mig	LHLH
ta.lab.'tu.kum	LHLH

References

Anani, M. (1989). Incorrect stress placement in the case of Arab learners of English. IRAL- International Review of Applied Linguistics in Language Teaching, 27(1), 15–22.

Archibald, J. (1993). Language Learnability and L2 Phonology. Dordrecht: Kluwer Academic Publishers.

Antoniou, M., Ettlinger, M., & Wong, P. C. M. (2016). Complexity, Training Paradigm Design, and the Contribution of Memory Subsystems to Grammar Learning. *PLOS ONE*, *11*(7), e0158812.

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, *122*(3), 292–305.

Baptista, B.O. (1989). Strategies for the Prediction of English Word Stress. *International Review of Applied Linguistics* 27, 1, Feb,,1-14.

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Burzio, L. (1994). *Principles of English stress*. Cambridge University Press.

Chin, S. L. (2009). *The application of the less is more hypothesis in foreign language learning*. Florida Atlantic University Boca Raton, Florida.

Clopper, C. G. (2002). Frequency of stress patterns in English: A computational analysis. *IULC Working Papers Online*, *2*(2).

Cochran, B. P., McDonald, J. L., & Parault, S. J. (1999). Too Smart for Their Own Good: The Disadvantage of a Superior Processing Capacity for Adult Language Learners. *Journal of Memory and Language*, *41*(1), 30–58.

Conway, C. M., Ellefson, M. R., & Christiansen, M. H. (2003). When less is less and when less is more: Starting small with staged input. In *Proceedings of the 25th annual conference of the cognitive science society* (pp. 270–275). Lawrence Erlbaum Mahwah, NJ.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech & Language*, *2*(3–4), 133–142.

Demuth, K. (1993). Issues in the acquisition of the Sesotho tonal system. *Journal of Child Language*, *20*(02), 275-301.

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, *48*(1), 71–99.

Gibbs, S. (2004). Phonological awareness: an investigation into the developmental role of vocabulary and short-term memory. *Educational Psychology*, *24*(1), 13–25.

Hayes, B. (1995). *Metrical stress theory: Principles and case studies*. University of Chicago Press.

Hyman, L. M., & Kisseberth, C. (1998). Theoretical aspects of Bantu tone. CSLI.

Kapatsinski, V., Olejarczuk, P., & Redford, M. A. (under preparation). Perceptual learning of intonation: Adults are more narrow-minded than children. *Cognitive Science*.

Kersten, A. W., & Earles, J. L. (2001). Less Really Is More for Adults Learning a Miniature Artificial Language. *Journal of Memory and Language*, *44*(2), 250–273.

Kilany, Hanaa, et al. *Egyptian Colloquial Arabic Lexicon LDC99L22.* Web Download. Philadelphia: Linguistic Data Consortium, 1997.

Krashen, S. D., Long, M. A., & Scarcella, R. C. (1979). Age, Rate and Eventual Attainment in Second Language Acquisition. *TESOL Quarterly*, *13*(4), 573–582.

Lai, J., & Poletiek, F. H. (2010). The impact of starting small on the learnability of recursion. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Lai, J., & Poletiek, F. H. (2011). The impact of adjacent-dependencies and staged-input on the learnability of center-embedded hierarchical structures. *Cognition*, *118*(2), 265–273.

Lai, J., & Poletiek, F. H. (2013). How “small” is “starting small” for learning hierarchical centre-embedded structures? *Journal of Cognitive Psychology*, *25*(4), 423–435.

Lassiter, D. (2014) Submiterator. Retrieved from http://github.com/danlassiter/Submiterator

Ludden, D., & Gupta, P. (2000). Zen in the art of language acquisition: Statistical learning and the less is more hypothesis. In *22nd Annual Conference of the Cognitive Science Society*. Citeseer.

Newport, E. L. (1990). Maturational constraints on language learning. Cognitive Science, 14, 11–28.

Rohde, D. L. T., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: how important is starting small? *Cognition*, *72*(1), 67–109.

Schembri, T., Johnson, M., & Demuth, K. (2016a). Online learning of Cairene Arabic word stress patterns over time.

Schembri, T., Johnson, M., & Demuth, K. (2016b) Longi. A Simple Automated System for Conducting Longitudinal Studies on Amazon Mechanical Turk

Sereno, J. A. (1986). Stress pattern differentiation of form class in English. *The Journal of the Acoustical Society of America*, *79*(S1), S36–S36.

Siegelman, N., & Arnon, I. (2015). The advantage of starting big: Learning from unsegmented input facilitates mastery of grammatical gender in an artificial language. *Journal of Memory and Language*, *85*, 60–75.

Snow, C. E., & Hoefnagel-Höhle, M. (1978). The Critical Period for Language Acquisition: Evidence from Second Language Learning. *Child Development*, *49*(4), 1114–1128.

Taylor, P., Black, A. W., & Caley, R. (1998). The Architecture Of The Festival Speech Synthesis System. In *IN THE THIRD ESCA WORKSHOP IN SPEECH SYNTHESIS* (pp. 147–151).

Chapter 5: Conclusion

Summary of Results

The research contained in this thesis contributes to our understanding of the acquisition of main stress systems by exploring the second language (L2) acquisition of the Cairene Arabic stress system. The aim of the research was to quantify the factors underpinning successful acquisition: both linguistic and non-linguistic. Chapter 3 aimed to identify a number of linguistic factors affecting acquisition, finding that L1 transfer and overgeneralization had a significant effect on participants' performance. Chapter 4 demonstrated that presentation order had an important effect on successful acquisition, such that participants who were immediately presented with the full grammatical complexity of the stress system were better able to acquire it. In addition to these findings, this thesis introduces and implements a set of methodological tools and ideas intended to facilitate future research. Chapter 2 introduced Longi, a piece of software designed to facilitate the use of longitudinal, or multi-session, experiments in Mechanical Turk. This software enabled the research conducted in later chapters, as the complex Cairene Arabic stress system could not have been acquired in a single session. More generally, this software is intended to enable future acquisition research, particularly in the area of understudied languages.

In Chapter 3, adult participants were taught the Cairene Arabic stress system over the course of 4 days. Artificial language learning (ALL) methodology was combined with real-language data extracted from the LDC Colloquial Egyptian Arabic lexicon (Kilany et al., 1997). Optimality Theory (OT) constraints were used to group the stimuli into groups of similar words. Participants were presented with stimuli of gradually increasing complexity. This was expected to make acquisition easier, due to the predictions of the ‘less is more’ hypothesis (Newport, 1990). This states that children’s limited working memory capacity helps them during the course of L1 acquisition. Therefore, acquisition in adults can be improved by simulating limited working memory during acquisition. One method of doing this is to limit the input initially provided in terms of grammatical complexity. Participants consistently scored an average of 80% correct in the training phase across all days. Additionally, they scored between 64% and 70% in the generalization phase. This was comparable to Carpenter’s (2005, 2010) studies, in which a simpler stress system was acquired. In Carpenter’s studies, participants averaged 89-90% correct in training, and 61-70% in generalization, depending on condition. Scores were higher at the beginning of the experiment, dropping as stimuli gradually increased in complexity. As OT constraints were used to group the stimuli, an attempt was made to extract differential rates of acquisition for each set of constraint rankings, in order to determine whether certain constraint rankings were easier or more difficult to acquire than others. However, this proved to be a difficult task for a number of reasons, discussed in greater detail below.

Due to these methodological challenges, an alternate analysis, based on the similarities and differences between the English and Cairene Arabic stress systems, was carried out. Based on this analysis, it was discovered that participants were highly sensitive to the position of main stress in the word. Participants performed best on words with medial stress, with no difference in second- or third- syllable stress in the case of 4-syllable words. Participants performed worst on words with initial stress. Participants’ poor performance on words with initial stress was surprising. In these contexts, both the L1 and target language predict initial stress. This should make these words easier to acquire. This result was hypothesized to be an overgeneralization effect, which is a common pattern seen in the L2 acquisition of main word stress (Guion, Harada, & Clark, 2004; Lord, 2001).

In addition, differential rates of acquisition were demonstrated for two sets of constraint rankings. Participants performed significantly better on words in the constraint group WSPμμμ, which implies the ranking WSPμμμ >> NonFin, compared to words in the constraint group WSP-CVV, which implies the ranking WSPμμμ >> NonFin. Finally, participants showed no evidence of sensitivity to the effect of weight independent of the presence of a long vowel. In other words, the presence of a CVV syllable triggered stress assignment, while the presence of a CVC syllable did not. This was explained as an L1 transfer effect, as the results mirror participants' stress assignment strategies in English novel word experiments (Domahs, Plag, & Carroll, 2014; Guion, 2005; Guion, Clark, Harada, & Wayland, 2003; Guion et al., 2004).

The experiment in Chapter 4, tests the 'less is more' hypothesis (Newport, 1990). In this experiment, the stimuli, presentation and methodology of Chapter 3 were replicated. However, stimuli were presented in random order, such that words of any level of complexity could be presented at any point in the experiment. The 'less is more' hypothesis predicts that participants in the random experiment (Chapter 4) should demonstrate worse performance than those in the staged input experiment (Chapter 3), in which input was initially limited in terms of grammatical complexity. By the end of day 3, participants in both experiments had been exposed to the same set of stimuli, but in different orders. On day 4, participants in both experiments were tested on the same novel set of stimuli, which represented the full range of grammatical variation introduced on days 1-3. This facilitated a direct comparison between participants in the two experiments, making performance on day 4 the most significant indicator of the way in which presentation order affects performance. Overall performance was quite different for participants across the two experiments. Participants in the random experiment performed especially well on words with final superheavy syllables, and performed particularly poorly on words with initial stress or stressed light syllables.

The overall trajectory of performance across the two experiments clearly demonstrated that participants in the random experiment (Chapter 4) outperformed those in the staged input experiment (Chapter 3). In the staged input experiment, participants' performance was best on the first day of the experiment, adjusted for other factors, and dropped thereafter. In the random experiment, participants' performance continued to improve throughout the experiment. Importantly, in the random experiment, participants' performance on day 4 was significantly better than that of participants in the staged input experiment.

Both sets of participants had some patterns of acquisition in common. The presence of a long vowel had a significant effect on the stress assignment strategies in both studies, independent of the effect of weight, indicating that participants had not fully acquired the underlying Cairene Arabic weight system. Additionally, both sets of participants performed equally poorly on words with initial stress, an unexpected result which was hypothesized to be due to overgeneralization. Finally, both sets of participants performed better on words in the constraint group WSPμμμ compared to words in the constraint group WSP-CVV. However, participants in the random experiment outperformed those in the staged input experiment on a number of metrics measuring their knowledge of aspects of the underlying prosodic structure of the language. Participants in the random experiment performed better on words with final stress, which is uncommon in English nouns, demonstrating a lessened L1 transfer effect. Additionally, they acquired a better understanding of the weight system, such that they were more likely to stress heavy (CVC, CVV) syllables over light (CV) syllables.

Limitations of the Studies

One major limitation of this research is that it was not possible to determine to what extent participants acquired the full set of constraint rankings underlying the Cairene Arabic stress system. This is in large part because sets of constraint rankings coincided with surface properties of the language, making it difficult to determine whether participants were learning the rankings themselves, or whether they were simply sensitive to these surface properties. For example, in Chapter 3, participants performed particularly well on the following sets of constraints: EndRule & WSP, NonFin & EndRule, NonFin & EndRule & WSP, and NonFin & WSP. All the words in these constraint groups contained a heavy penultimate syllable. This made it difficult to state with confidence that participants had acquired the constraint rankings in question. Similarly, in Chapter 4, participants performed particularly poorly on the following sets of constraints: NonFin, Trochaic & A-L, Trochaic & A-L & EndRule, Trochaic & A-L & EndRule & NonFin, and Trochaic & A-L & NonFin. All of the words in four of these constraint groups contained only stressed light syllables, while the majority of words in the remaining constraint group, NonFin, contained stressed light syllables. Because of these distributional properties, it is difficult to determine whether participants had acquired these constraint rankings, or whether they had acquired a general prohibition on stressing light syllables. Additionally, there were inherent contradictions in hypothesizing that participants had acquired the constraint rankings in Chapter 3. For example, if one assumes that participants had acquired the constraint rankings in the constraint group NonFin & EndRule-R & WSP, this would necessarily imply that the three individual active constraints, NonFin, EndRule-R and WSP had all been acquired and ranked appropriately. However, participants' performance on words in the constraint group NonFin was significantly worse than on any other constraint group. This suggests that the hypothesis that participants have acquired the constraint groups in question is incorrect. The results for Chapter 4 do not contain any such contradictions: participants performed poorly on all constraint groups containing any combination of the active constraints ALIGN-L, TROCHAIC or NONFIN. This is an encouraging result. Nevertheless, given the contradictory results in Chapter 3, and the fact that a simpler explanation for participants' patterns of acquisition exists, there is not enough evidence to conclude that it is constraint rankings, rather than the presence of stressed light syllables, which is driving participants' performance in Chapter 4. Another important note is that the constraint set used to describe the Cairene Arabic stress system is by no means the only possible OT analysis of these patterns. The adoption of an alternate set of OT constraints may lead to different conclusions about participants' acquisition of constraint rankings vs surface stress patterns. One alternate analysis involves dropping WSP-CVV and WSPμμμ from the constraint set. The pattern of final stress in CVV, CVVC and CVCC – but not CVC – syllables could then be derived by assuming that final consonants in Cairene Arabic do not project moras. Under this analysis, the patterns of final stress, as well as the assignment of stress to non-final CVC syllables, would be governed by the constraints WSP and FINAL-C. However, participants' variable performance in these contexts suggests that this analysis may not reflect the constraints acquired throughout the experiment. In particular, this analysis would predict no difference in performance between CVV-final and CVVC-final syllables. Yet participants do perform significantly differently on words containing these syllables. This suggests that the constraint set used throughout this dissertation is a more accurate representation of participants' acquisition.

Few studies have directly attempted to demonstrate L2 acquisition of specific OT constraint rankings in an experimental setting (Bahl, Plante, & Gerken, 2009; Gerken, 2004; Guest, Dell, & Cole, 2000; Plante, Bahl, Vance, & Gerken, 2010). These studies all use the same set of stimuli and constraints. A closer examination of the artificial languages used in these studies demonstrates how the limitations of the current research may be overcome in future studies, yet also illustrates further challenges. The artificial languages in these studies obey a simple set of four constraints (Guest et al., 2000). The constraint CLASH (A) states that two adjacent syllables cannot both be stressed; the constraint HEAVY (B) states that heavy syllables must be stressed; the constraint PENULT (C) states that all penultimate syllables must be stressed; the constraint LEFT-ALTERNATING (D) states that stress must be assigned to every odd-numbered syllable, beginning with the first syllable. Stimuli were designed to induce conflicts between constraints. For example, if a word contains a heavy syllable adjacent to the penultimate syllable, it is not possible to satisfy all constraints: stressing both the heavy and penultimate syllables violates CLASH. Each word presented to participants demonstrated a single constraint ranking. Participants were presented with evidence for three sets of rankings: A >> B; B >> C; C >> D. Crucially, no evidence was presented for a fourth ranking: B >> D. However, if participants had truly acquired the three rankings, they should be able to correctly induce the fourth ranking through the principle of transitivity, even without direct evidence. At the end of the experiment, participants were tested on unseen words which placed constraint B and D in direct conflict for the first time.

This methodology provides a way to conclusively determine whether participants are merely sensitive to surface stress patterns, or truly acquiring the underlying constraints. If participants consistently induce the ranking B >> D without being exposed to any stimuli which exemplify this ranking, this is strong evidence that the three constraint rankings presented to them were truly acquired. However, Guest et al. (2000) found only weak evidence that participants had acquired the ranking B >> D. Bahl et al. (2009) replicated this result, finding that adult participants were unable to induce a constraint ranking through transitivity. However, participants were able to successfully induce this ranking when stimuli were manipulated such that the acoustic realization of stress was greatly emphasized and unusually salient. Conversely, Gerken (2004), following a similar methodology, found that 9-month-old infants were able to consistently induce a constraint ranking through transitivity without any special manipulations.

The above set of studies demonstrate that it is possible to conclusively determine that participants have acquired a set of constraint rankings in an experimental setting. However, this method cannot be easily applied to a setting in which stimuli are drawn from a complex stress system, rather than an artificial language. The artificial language used in these studies is much simpler than the Cairene Arabic stress system. There are only four constraints, and all of these constraints are ranked with respect to each other in a straightforward manner. There are no two constraints whose relative ranking is unknown. Each stimulus item demonstrates a single constraint ranking. Real language data is considerably messier. There are nine active constraints in the OT analysis of Cairene Arabic contained in Chapter 3. Existing lexical items provide insufficient evidence of a ranking between a number of these constraints. For example, this analysis does not rank the constraints ENDRULE-R and NONFIN with respect to each other. While both constraints are active in determining the stress patterns of a number of Cairene Arabic words, they never come into direct conflict with one another. The constraint ENDRULE-R states that main stress must fall within the rightmost foot in the word. In other words, it is satisfied if either the first or second syllable in this foot receives main stress. The constraint NONFIN prohibits main stress on the final syllable. As Cairene Arabic is a trochaic stress system, main stress falls on the first syllable of a foot. Therefore, lexical items containing two syllables in the rightmost foot are able to satisfy both ENDRULE-R and NONFIN. A foot which is binary in terms of moras may contain two light syllables or a single heavy syllable. If a word contains a final heavy syllable, this might bring about a conflict between ENDRULE-R and NONFIN. However, NONFIN ensures that this final syllable remains unfooted, satisfying both

ENDRULE-R and NONFIN. Finally, NONFIN is violated in lexical items which contain final stress, but this is due to the constraints WSPμμμ and WSP-CVV – the constraint ENDRULE-R does not come into play. Therefore, there is no direct surface evidence of the ranking between these two constraints. Further, while stimuli in artificial language experiments are designed to demonstrate a ranking between two individual constraints, many lexical items in the current experiments require a larger set of active constraints in order to correctly predict stress assignment. An additional challenge lies in the well-known problem of hidden structure (Apoussidou, 2007; Jarosz, 2009; Tesar, 2004). The correct construction of metrical feet is necessary to derive correct stress assignment in Cairene Arabic. However, the location of these feet is not directly evident in the input available to the learner, who must first acquire the grammar in order to be able to construct feet accurately. Similarly, evidence of learners' foot construction, whether correct or incorrect, is not directly available to an outside observer. Acquisition of the constraints TROCHAIC and ALIGN-L is necessary in order to fully acquire the Cairene Arabic stress system, yet it is unclear how to test whether these constraints have been ranked appropriately. It is not immediately clear how to resolve the above issues in order to determine acquisition rates for the full set of constraints active in Cairene Arabic stress assignment.

Another limitation of the current studies lies in their experiment design. The stimuli in the experiment range from 2-5 syllables long. Because of this, when a participant is presented with a correct and incorrect choice, the incorrect item is selected from a set of multiple possible incorrect items. For example, in the case of a 4-syllable word with final stress, participants could potentially be presented with an incorrect alternative with a) initial stress; b) stress on the second syllable; c) stress on the third syllable. 2-syllable words have only one possible incorrect alternative, while 5-syllable words have 4 possible incorrect alternatives, and so on. Each incorrect alternative presented to participants is selected at random from the total set of possible alternatives. It is possible that the choice of incorrect alternative may have some effect on participants' performance. For example, there is no significant difference between participants' performance on medial syllables, regardless of whether they occur as the 2nd or 3rd syllable. Therefore, given a 4-syllable word with medial stress on the 2nd syllable, participants may find it particularly difficult to answer correctly when the incorrect alternative presented contains medial stress on the 3rd syllable. On the other hand, when correct and incorrect alternatives are maximally different, this may aid acquisition. For example, given a 4-syllable word with final stress, participants may find it particularly easy to answer correctly when an incorrect alternative with initial stress is presented.

Although it is easy to imagine that such considerations may have some effect on acquisition, the current experiment design makes it impossible to check for these effects. The experiment software used for the current experiments does keep track of the incorrect alternative presented to participants. A naïve way to tackle this question would be to include 'Incorrect_Alternative' as a factor in one of the statistical models presented in this dissertation. For example, the formula for the surface stress model in Chapters 3 and 4 is Correct ~ Stress_Position + Weight + Long_Vowel + Day + (1 | Worker_ID) + (1 | Item_Number). This could be amended to Correct ~ Stress_Position + Weight + Long_Vowel + Day + **Incorrect_Alternative** + (1 | Worker_ID) + (1 | Item_Number). In this case, the factor Incorrect_Alternative would state whether the incorrect alternative presented to participants contained stress on the first, second, third, fourth or fifth syllable. However, this does not adequately address the question at hand. An incorrect alternative with stress on the 2nd syllable may have very different effects depending on the other properties of the item in question. For example, given a 2-syllable word, this incorrect alternative is the only option; given a 4-syllable word with medial stress on the 3rd syllable, this incorrect alternative is maximally similar to the correct choice; given a 4-syllable word with final stress, this incorrect alternative is distinctly different from the correct choice. These considerations are complicated further when taking weight into account. For example, given a 4-syllable word, it may be easier to distinguish between correct medial stress on the 2nd syllable, and incorrect medial stress on the 3rd syllable when the correct syllable is heavy and the incorrect syllable is light. Given these considerations, it is clear that this question cannot be addressed by simply adding Incorrect_Alternative as a factor. Instead, a model which aims to determine whether there is any effect of the selected incorrect alternative on performance must include an interaction between each individual item and the incorrect alternative presented to participants. The formula for such a model would be as follows: Correct ~ Stress_Position + Weight + Long_Vowel + Day + **Incorrect_Alternative * Item_Number** + (1 | Worker_ID) + (1 | Item_Number). There are 195 individual items included in the current experiment. The average word length of stimuli is 3.27 syllables, rounded down to 3 for convenience. Therefore, there is an average of 2 incorrect alternatives for each item. This means that, for the interaction factor Incorrect_Alternative * Item_Number alone, there are 2 * 195 = 390 separate effects to be included in such a model. Table A5 in Chapter 4 demonstrates that, given the dataset compiled for these experiments, there was insufficient data for a model with a total of 25 effects to produce interpretable results. Observing the disparity between these numbers, it is clear that there is insufficient data to adequately explore this question given the current experiment design and the amount of data collected.

There are a number of different ways in which future experiments could be designed to tackle this question. Theoretically, it would be possible to dramatically increase the number of participants included in the study, thus increasing the amount of data collected. However, this approach is infeasible due to concerns of time and cost. Another possibility would be to drastically reduce the number of items presented to participants, in order to cut down the number of interactions to a more manageable figure. However, this would change the overall nature of the experiment. If participants are presented with only a small set of stimuli, they cannot acquire a complex system such as the Cairene Arabic stress system. A final possibility would be to conduct an analysis of the Cairene Arabic stress system in terms of a model which ranks correct and incorrect candidates in terms of grammaticality. For example, Harmonic Grammar (Potts, Pater, Jesney, Bhatt, & Becker; 2010) assigns a score to each possible output by adding weighted constraint violations. The correct candidate is the one with the highest score. However, this system makes it possible to compare the grammaticality of incorrect candidates. For example, a 4-syllable word with final stress scoring -1 may have an incorrect alternative with initial stress scoring -10; with stress on the second syllable scoring -5; with stress on the third syllable scoring -6. In this hypothetical example, it is easy to see that the incorrect alterative with initial stress is the least grammatical, while there is little difference between stress on the second or third syllable. These harmony scores would be computed separately for each individual item. Any surface patterns which affect grammaticality would have the effect of assigning violation marks to candidates, and therefore be reflected in the harmony score. The harmony score is therefore a sum of all factors affecting an item’s grammaticality, including weight and contextual positional effects. Because of this, it would not be necessary to compute interactions for each individual items. Instead, a simple model including the factor ‘Harmonic_Score’ would do the work of the interaction ‘Incorrect_Alternative * Item_Number’. This would greatly reduce the complexity of the resulting statistical model. This analysis could be carried out on the current dataset that has already been collected. However, this would require a Harmonic Grammar analysis of Cairene Arabic to be carried out. In addition, it would be necessary to calculate the Harmony Score of each individual item, and all its incorrect alternatives, in the current experiment. This is not a trivial task, requiring computational analysis. As such, this is outside the scope of the current study, but remains an open question for future research.

Future Directions

This section outlines a number of future extensions and applications of the research contained in this thesis. Chapter 2 introduces Longi, a piece of software designed for use in this thesis, which has many applications in linguistics as well as other fields. Conducting learning studies over multiple days can aid acquisition. Increasing the time between learning sessions improves retention of taught material (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006), and sleep consolidation has an additional positive effect on performance (Fenn, Margoliash, & Nusbaum, 2013). Because of this, participants in multi-day acquisition studies are able to acquire more complex structures, and retain more of what they are taught, than those participating in acquisition studies held over a single day. Longi allows such studies to be held online, considerably reducing the time and cost involved. This may enable studies into the acquisition of complex interacting structures which were previously impractical to study experimentally. For example, long-term studies could replicate the classroom environment over an extended period of time, in which participants are taught a wide range of linguistic structures in a given language. In Chapter 3, the use of Longi is combined with Artificial Language Learning (ALL) methodology and real-language input, demonstrating how the use of technology can enable research into the L2 acquisition of understudied languages. Future research may apply this design to the acquisition of a wider range of languages and linguistic structures.

The results of the experiments described in Chapters 3 and 4 raise a number of unanswered questions which would be interesting to explore in future research. In Chapter 3, participants' performance was best on the first day of the experiment, but then dropped as the stimuli presented to participants increased in complexity. However, in Chapter 4, participants' performance had not reached a peak by the end of the experiment. Performance was still improving linearly by the end of day 3, when training ended. This suggests that if this experiment was replicated with a greater number of stimuli, and run over a longer period of time, participants' performance would continue to improve. There are a number of possible scenarios. Participants may gain a better understanding of the Cairene Arabic weight system, learning that long vowels and coda consonants should have an equal effect on stress assignment. If this were the case, the factor Long Vowel would no longer have an effect on participants' performance independent of the effect of Weight. Alternatively, participants may learn that, although heavy syllables are preferentially stressed over light syllables in most contexts, there exist exceptional patterns in which light syllables are stressed over heavy syllables. Finally, participants may perform better on words with initial stress, demonstrating a lessened overgeneralization effect. Conversely, participants' performance may plateau, showing no further improvement.

A common theme throughout this dissertation was that participants' native stress patterns significantly affected their performance. For example, participants' performance in Chapter 3 was affected by syllable weight in a manner which mirrored English, rather than Cairene Arabic, stress patterns. L1 transfer effects were also observed for participants' performance in terms of stress position. Given these results, it may be interesting to replicate these experiments with native speakers of a range of different L1s, using the same stimuli and methodology. Statistical analysis of the differences in performance between speakers of a range of L1s would enable us to better quantify which patterns of acquisition are attributable to L1 transfer effects, and which are common to all participants. English has a probabilistic stress system. Therefore, English participants may be predisposed to search for probabilistic, rather than deterministic, patterns when exposed to an L2 stress system. In contrast, speakers of stress systems in which stress assignment is variable yet deterministic may be predisposed to search for the kinds of patterns found in Cairene Arabic. It would therefore be interesting to compare their performance against that of English participants.

In Chapter 4, participants were shown to perform better when input was ordered randomly rather than presented in order of complexity. Stimuli were ordered in terms of complexity according to the set of active OT constraints required to correctly assign stress. However, there are a number of logically possible ways of sorting stimuli by complexity. For example, stimuli could be sorted by length, such that participants are presented only with 2-syllable words on day 1, with 3-syllable words on day 2, and so on. If it is true that limiting the presentation of longer words harms participants' ability to form correct hypotheses by ruling out competing hypotheses, participants' performance would be worse than in either Chapter 3 or 4. Similarly, manipulating the organization of stimuli in various ways may have interesting effects on participants' patterns of acquisition. In Chapter 4, we hypothesized that the 'less is more' hypothesis may not apply equally to the acquisition of all linguistic domains. Stress and tone systems are somewhat unusual in that longer words provide significantly more information about the system than do shorter words. This may be the reason why we found the inverse of a "less is more" effect in Chapter 4. This hypothesis can be tested by comparing participants' performance when stimuli are initially limited versus randomly presented, across a wider range of stress, tone and other linguistic systems. If a "less is more" effect is found for the acquisition of phonological, morphological or syntactical systems, but not for the acquisition of tone or stress, this would be evidence that these systems are acquired differently due to their fundamental structure. These results have broad implications for L2 language instruction in the classroom. Learners may be better able to acquire certain structures, such as recursion, when input is initially limited. However, other structures, such as tone or stress, may be better acquired when the full complexity of the language is immediately available. A greater understanding of the conditions under which 'less is more' may inform pedagogical practice and enable L2 learners to better acquire their target language.

The 'less is more' hypothesis aims to explain the differential patterns of acquisition between children and adults learning language. However, the majority of research into this area has studied only adult participants. To our knowledge, only one study in this area explicitly compares the performance of children and adults on the same set of stimuli, and using the same methodology and presentation. Kapatsinski, Olejarczuk, & Redford (2016) compared the acquisition of intonation contours in 3 sets of participants: children younger than 10 years old, children aged between 10 and 11, and adults. They found that the overall patterns of acquisition seen in younger children were different from those seen in older children and adults; however, in certain contexts, both sets of children patterned together compared to adults. These results demonstrate that conducting these experiments only on adults may result in incomplete information. Experimental data from children is also necessary to fully understand the patterns of acquisition at play. This suggests another potential avenue for future research. The experiments in Chapter 3 and 4 could be replicated with child participants, whose performance may shed further light on the processes underpinning acquisition. For example, such research may find that children also perform better with random rather than staged input; or that the performance boost for random input is greater or smaller in children than adults; or that children perform better with staged rather than random input. Each of these results would have different implications for our understanding of the children and adult's different patterns of acquisition in language learning. Kapatsinski et al. (2016) found different patterns of acquisition for younger and older children, suggesting that it may be useful to test children of a range of different ages. The two-forced choice paradigm used in Chapters 3 and

4 has been replicated on the Apple iPad, and has been used to test children as young as 3 (Xu Rattanasone, Davies, Schembri, Andronos, & Demuth, 2016). Artificial stress systems have been taught to children as young as 9 months old using eye tracking technology (Gerken, 2004; Saffran & Thiessen, 2003). The use of these technologies can enable further research into the "less is more" hypothesis, tracking the performance of participants of a variety of ages on the same set of stimuli with either initially limited or random presentation.

The 'less is more' hypothesis has been explored computationally as well as experimentally. Newport's (1990) early reports on the 'less is more' hypothesis were confirmed by computational models run by Elman (1993), who showed that a neural network was better able to process complex sentences when it was initially presented with shorter strings of language. In order to further test these results, another experiment was conducted. This time, the network was presented with full sentences throughout, but began training with severe memory limitations, which were gradually lifted. This network was also able to outperform one which began training with full sentences and no memory limitations. This result provided computational support for the 'less is more' hypothesis. Conversely, Rohde & Plaut (1999) attempted to replicate Elman's results, but found no performance advantage for a network which began training on limited input, or for a network which began training with memory limitations. This research was presented as a refutation of Elman's results. Computational models are especially useful when they are trained on the same set of stimuli, presented in the same order, as human participants. This facilitates a direct comparison between experimental and computational research. If both human participants and computational models perform similarly, this is strong evidence that the patterns of acquisition demonstrated are robust. Arnon & Ramscar (2012) constructed two computational models which were trained on the same stimuli as human participants, arranged in the same number of trials and blocks as those presented to participants. The first model was first trained on nouns, then full sentences containing those nouns. In the second model, the order of training was reversed. Similarly, two sets of human participants were trained on these two conditions. Both human participants and computational models acquired article-noun pairings significantly better when full sentences were presented first. As both sets of results were in agreement, this was taken as strong evidence that the 'less is more' effect does not hold for the acquisition of article-noun pairings. Similarly, the experimental results contained in this thesis could be strengthened through the use of computational modelling. If a computational model, trained on the same set of stimuli as participants in Chapter 3 and 4, performed better on random input as compared to initially limited input, this would further strengthen the hypothesis that the 'less is more' effect does not apply to the acquisition of main word stress. Additionally, such simulations could be run on a wide variety of stress and tone systems, further testing the hypothesis that these kinds of systems are better acquired when the full range of complexity is immediately available in the input.

Computational modelling may also shed further light on the question of whether participants are truly acquiring OT constraint rankings, or whether they are simply sensitive to surface properties of the input. This problem is discussed in greater detail in the section above. A variety of computational models are available in which acquisition of a linguistic system is based entirely on finding the correct ranking of OT constraints (Berent, Wilson, Marcus, & Bemis, 2012; Goldwater & Johnson, 2003; Hayes & Wilson, 2008). These models begin training with knowledge of a set of OT constraints. They receive as input a set of words, as well as any violation marks incurred by each word for each constraint. A number of additional models have been proposed specifically for the acquisition of metrical stress (Daelemans, Gillis, & Durieux, 1994; Gupta & Touretzky, 1992, 1994; Heinz, 2009). These models differ considerably in their internal architecture, and take in strings of light and heavy syllables as input. These models can be tested on the same set of stimuli as participants in Chapters 3 and 4. If the output of computational models based on surface properties of the string is more similar to participants' performance than those based on the ranking of OT constraints, this would provide supporting evidence that participants in these experiments may be more sensitive to surface properties such as stress position and weight. In addition, comparisons between the performance of human participants and computational models may provide further insights into which aspects of the Cairene Arabic stress system are inherently harder to acquire. For example, the participants in Chapter 4 performed poorly on words which contain light stressed syllables. Computational models may take a longer amount of time to acquire these words relative to other stimuli, or fail to acquire them altogether. If this is the case, this would suggest that words containing light stressed syllables are relatively difficult to acquire within the context of the Cairene Arabic stress system. Conversely, there may be no correlation between the types of words which human participants and computational models find relatively difficult to acquire. This would suggest that human participants are using learning strategies different to those simulated in computational learners. Such comparisons may be complicated by the fact that, while human participants will always have prior knowledge of an L1, computational models do not. As such, patterns of acquisition attributable to L1 transfer would not be observed in computational models. One solution may be to train computational models on large datasets corresponding to the human participants' L1 prior to beginning the experiment. However, to our knowledge, this approach has not been tested, and may result in an inability to successfully acquire the L2. Even if this is not the case, computational models are currently unable to achieve a level of L1 competence comparable to that of a human native speaker. Therefore, computational models are likely to serve only as an imperfect analogue of human performance in the foreseeable future.

The research contained in this thesis has a number of broad implications and applications. The experimental results in Chapters 3 and 4 raise a number of questions in the field of L1 and L2 acquisition. The extent of participants' ability to fully acquire the Cairene Arabic stress system is unknown, as performance had not yet plateaued by the end of the experiment in Chapter 4. More generally, a number of questions regarding the nature of maturational constraints on language acquisition remain unanswered. The literature on the 'less is more' hypothesis suggests that this effect applies to some linguistic domains but not others. However, further research is required to determine the conditions under which a 'less is more' effect can be demonstrated. These results have important implications for L2 instruction in the classroom. Finally, the introduction of Longi, which enables researchers to carry out online longitudinal research more easily, has broad implications for acquisition research, as well as for researchers carrying out online experimentation in other fields.

References

Apoussidou, D. (2007). *The learnability of metrical phonology*. LOT.

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, *122*(3), 292–305.

Bahl, M., Plante, E., & Gerken, L. (2009). Processing prosodic structure by adults with language-based learning disability. *Journal of Communication Disorders*, *42*(5), 313–323.

Berent, I., Wilson, C., Marcus, G. F., & Bemis, D. K. (2012). On the Role of Variables in Phonology: Remarks on Hayes and Wilson 2008. *Linguistic Inquiry*, *43*(1), 97–119.

Carpenter, A. C. (2005). Acquisition of a natural vs. unnatural stress system. In *Proceedings of the 29th annual Boston University Conference on Language Development* (pp. 134–43).

Carpenter, A. C. (2010). A naturalness bias in learning stress. *Phonology*, *27*(3), 345–392.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354.

Daelemans, W., Gillis, S., & Durieux, G. (1994). The acquisition of stress: A data-oriented approach. *Computational Linguistics*, *20*(3), 421–451.

Domahs, U., Plag, I., & Carroll, R. (2014). Word stress assignment in German, English and Dutch: Quantity-sensitivity and extrametricality revisited. *The Journal of Comparative Germanic Linguistics*, *17*(1), 59–96.

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, *48*(1), 71–99.

Fenn, K. M., Margoliash, D., & Nusbaum, H. C. (2013). Sleep restores loss of generalized but not rote learning of synthetic speech. *Cognition*, *128*(3), 280–286.

Gaja Jarosz. (2009). Naive Parameter Learning for Optimality Theory - The Hidden Structure Problem. Presented at the 40th Annual Meeting of the North East Linguistic Society, MIT Cambridge MA.

Gerken, L. (2004). Nine-month-olds extract structural principles required for natural language. *Cognition*, *93*(3), B89–B96.

Goldwater, S., & Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm workshop on variation within Optimality Theory* (pp. 111–120).

Guest, D. J., Dell, G. S., & Cole, J. S. (2000). Violable Constraints in Language Production: Testing the Transitivity Assumption of Optimality Theory. *Journal of Memory and Language*, *42*(2), 272–299.

Guion, S. G. (2005). KNOWLEDGE OF ENGLISH WORD STRESS PATTERNS IN EARLY AND LATE KOREAN-ENGLISH BILINGUALS. *Studies in Second Language Acquisition*, *27*(4), 503–533.

Guion, S. G., Clark, J. J., Harada, T., & Wayland, R. P. (2003). Factors Affecting Stress Placement for English Nonwords include Syllabic Structure, Lexical Class, and Stress Patterns of Phonologically Similar Words. *Language and Speech*, *46*(4), 403–426.

Guion, S. G., Harada, T., & Clark, J. J. (2004). Early and late Spanish–English bilinguals' acquisition of English word stress patterns. *Bilingualism: Language and Cognition*, *7*(3), 207–226.

Gupta, P., & Touretzky, D. S. (1992). A Connectionist Learning Approach to Analyzing Linguistic Stress. In I. J. E. Moody, S. J. Hanson, & R. P. Lippmann (Eds.) (Vol. 4).

Gupta, P., & Touretzky, D. S. (1994). Connectionist models and linguistic theory: Investigations of stress systems in language. *Cognitive Science*, *18*(1), 1–50.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*(3), 379–440.

Heinz, J. (2009). On the role of locality in learning stress patterns. *Phonology*, *26*(2), 303–351.

Kapatsinski, V., Olejarczuk, P., & Redford, M. A. (under preparation). Perceptual learning of intonation: Adults are more narrow-minded than children. *Cognitive Science*.

Kilany, Hanaa, et al. Egyptian Colloquial Arabic Lexicon LDC99L22. Web Download. Philadelphia: Linguistic Data Consortium, 1997.

Lord, G. E. (2001). *The second language acquisition of Spanish stress: Derivational, analogical or lexical?* (Ph.D.). The Pennsylvania State University, United States -- Pennsylvania.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*(1), 11–28.

Plante, E., Bahl, M., Vance, R., & Gerken, L. (2010). Children with specific language impairment show rapid, implicit learning of stress assignment rules. *Journal of Communication Disorders*, *43*(5), 397–406.

Rohde, D. L. T., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: how important is starting small? *Cognition*, *72*(1), 67–109.

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

Tesar, B. (2004). Using inconsistency detection to overcome structural ambiguity. *Linguistic Inquiry*, *35*(2), 219–253.

Xu Rattanasone, N., Davies, B., Schembri, T., Andronos, F., & Demuth, K. (2016). The iPad as a Research Tool for the Understanding of English Plurals by English, Chinese and Other L1 Speaking 3-and 4-year-olds. *Frontiers in Psychology*, *7*, 1773.