FREQUENCY EFFECTS ON ESL COMPOSITIONAL MULTI-WORD SEQUENCE PROCESSING

By

Sarut Supasiraprapa

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Second Language Studies – Doctor of Philosophy

2017

ABSTRACT

FREQUENCY EFFECTS ON ESL COMPOSITIONAL MULTI-WORD SEQUENCE PROCESSING

By

Sarut Supasiraprapa

The current study investigated whether adult native English speakers and English-as-a-second-language (ESL) learners exhibit sensitivity to compositional English multi-word sequences, which have a meaning derivable from word parts (e.g., *don't have to worry* as opposed to sequences like *He left the US for good*, where *for good* cannot be taken apart to derive its meaning). In the current study, a multi-word sequence specifically referred to a word sequence beyond the bigram (two-word) level. The investigation was motivated by usage-based approaches to language acquisition, which predict that first (L1) and second (L2) speakers should process more frequent compositional phrases faster than less frequent ones (e.g., Bybee, 2010; Ellis, 2002; Gries & Ellis, 2015). This prediction differs from the prediction in the mainstream generative linguistics theory, according to which frequency effects should be observed from the processing of items stored in the mental lexicon (i.e., bound morphemes, single words, and idioms), but not from compositional phrases (e.g., Prasada & Pinker, 1993; Prasada, Pinker, & Snyder, 1990). The present study constituted the first attempt to investigate frequency effects on multi-word sequences in both language comprehension and production in the same L1 and L2 speakers.

The study consisted of two experiments. In the first, participants completed a timed phrasal-decision task, in which they decided whether four-word target phrases were possible English word sequences. This task measured how fast participants receptively

process a phrase, with their reaction time being the outcome measure. In the second experiment, the same participants completed an oral elicitation task, in which they saw and orally produced target phrases. The outcome measure was the production durations of the first three words (e.g., *don't have to worry*) in the same target phrases used in the first experiment. Participants were a sample of native English speakers (N=51) and ESL learners (N=52) who can be characterized as being proficient enough to study in an English academic environment (mean internet-based TOEFL scores = 95.52, SD = 6.63) and who had lived in the US for 2-3 years (mean = 2.61 years, SD = 0.56).

The results from the first experiment suggested phrase frequency effects in both participant groups and countered the proposal that L2 learners cannot retain information about L2 word occurrences in their memory (Wray, 2002). These results support the prediction from usage-based approaches and further corroborate previous proposals that frequency data from large native English corpora should be representative of the regularities of English input that speakers in general are exposed to (Hoey, 2005; Wolter & Gyllstad, 2013). Moreover, the results entail a need for future L1 and L2 psycholinguistics model to accommodate phrase frequency effects. On the other hand, in the second experiment, both participant groups did not exhibit phrase frequency effects. In light of previous similar compositional multi-word sequence production studies (e.g., Bannard & Matthew, 2008; Ellis, Simpson–Vlach, & Maynard, 2008), which had yielded mixed results, and the results from the first experiment, the absence of the effects in the second experiment could have stemmed from cross-study methodological differences, including the type of experimental tasks used to investigate multi-word sequence frequency effects.

ACKNOWLEDGEMENTS

The completion of this dissertation was possible due to the support and help from many. First, my gratitude goes to my dissertation committee members. I am thankful to Charlene Polio for her guidance and for always having strong faith in me since I first entered this academic field in 2008. I am also grateful to Aline Godfroid for her detailed and insightful suggestions and for encouraging me to continue pursuing research in this area. My thanks also go to Shawn Loewen, who provided useful advice on the design of my study and statistical analyses. In addition, I thank Susan Gass for her suggestions and for her leadership of the Second Language Studies program, which has brought about several forms of support and opportunities for students, including a summer fellowship which allowed me to carry out this dissertation project in the summer of 2016.

Several other professors and friends also made significant contributions to this dissertation. Without Karthik Durvasula, with whom I took a phonology class as an MATESOL student, and Qian Luo from the linguistics department, the speech production experiment in this dissertation would have been a serious practical challenge. My thanks also go to Jens Schmidtke, who was always willing to provide suggestions about the R program no matter where he was, be it the US, Mozambique, Germany, or Jordan. Moreover, Nicole Jess at the MSU Center for Statistical Training and Consulting gave me useful advice about mixed effects regression modeling. I also thank Attakrit Leckcivilize for his statistical insight and the friendship we have formed since high school, as well as Suzanne Wagner and Suzanne Johnston for their respective suggestions on the Linguistics Data Consortium and Superlab. In addition, Stella He and several of my Thai

friends at MSU, including Pui, Kwan, Sorrachai, and Supawadee, helped me with participant recruitment.

I would also like to thank professors with whom I took classes during my time in the MATESOL and the SLS programs at MSU, including Cristina Schmitt, Debra Hardison, Patti Spinner, and Dianne Larsen–Freeman. The knowledge I gained from them has broadened my academic perspectives, increased my interest in linguistics and language acquisition, and influenced this dissertation, directly or indirectly, in various ways.

In addition, I am grateful to many good friends in Thailand for their encouragement and support during the past four years, including Torsak, Gaew, Kwang, Win, Seth, Sa, and Mon. Many people I befriended with at MSU also made my time in the US more fun and meaningful, particularly Ben, Se Hoon, Prare, Big, Yui, Kung, Tim, Pom, Tink, and Aoi. I also thank Sakol Suethanapornkul for his encouragement and inspiring academic determination. On some occasions, it was words of wisdom from these people that gave me the energy to go on.

Finally, I thank my parents, Saner Supasiraprapa and Weena Supasiraprapa, and my sister, Sukumal Supasiraprapa, for their unending love, patience, and support. I am also grateful to my uncle, Vithul Liwgasemsarn, for helping my mom and my sister in many ways when my dad passed away in 2013 and I had to be away from home to pursue this doctoral degree.

This dissertation was financially supported by the College of Arts and Letters with a Dissertation Completion Fellowship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Recent years have seen an increase in first language (L1) and second language (L2) research investigating human sensitivity to the statistical probabilities of linguistic patterns at various levels, ranging from morphemes to words, phrases, and syntactic patterns (e.g., Ambridge, Kidd, Rowland, & Theakston, 2015; Arnon, 2015; Ellis, 2002; Matthews, Lieven, Theakston, & Tomasello, 2005). A great deal of research in this area has been motivated by language acquisition models or approaches which posit that the creation and entrenchment of linguistic knowledge in a learner's mind is driven by experience or an accumulation of statistical probabilities of occurrence in previously-encountered linguistic input. Frequency of encounters is considered a key index of experience, and these theoretical frameworks include usage-based approaches to language acquisition (e.g., Bybee, 2010; Ellis, 2011, 2012; Goldberg, 1995, 2006; Gries & Ellis, 2015; Tomasello, 2003, 2009). From a usage-based perspective, linguistic representation involves various types of linguistic patterns with different degrees of complexity and abstraction, including words (e.g., *kick*), multi-word sequences (e.g., *He kicked the ball*), and more abstract constructions (e.g., *Subject-Verb-Object*). There is no complete separation of grammar and the mental lexicon, and words and multi-word sequences are represented by the same mechanism. In addition, speakers are predicted to demonstrate sensitivity to the distributional properties, particularly frequency, of not only single words, but also compositional word sequences—namely, those that have a meaning derivable from word parts (e.g., *He kicked the ball*).

This prediction has been borne out in a growing amount of empirical evidence demonstrating frequency effects on L1 comprehension and production in both children

and adults (for reviews see Ambridge et al., 2015; Arnon, 2015; Diessel, 2007; Ellis, 2002). Such evidence has significant implications for both the nature of linguistic representation and language processing models. The evidence is compatible with a linguistic model in which frequency influences the representation and processing of all linguistic patterns, not just individual words—that is, a model in which the strengthening of all linguistic patterns in a learners' representation result from frequency of previous encounters and the strengthening leads to an expectation of other elements in a pattern when speakers see an element in that pattern (e.g. Arnon, 2015; Jurafsky, 1996; McDonald & Shillcock, 2003; Jurafsky, Gregory & Raymond, 2000). Many scholars (e.g., Ambridge et al., 2015; Arnon & Snider, 2010; Arnon & Priva, 2013; Diessel, 2007) also pointed out that frequency effects on compositional phrases is not predicted by the traditional words-and-rules linguistic model, which argues for two distinct types of representations for words and for larger compositional phrases (e.g., Pinker & Ullman, 2002). In the traditional prediction, the processing of compositional sequences, which are computed based on grammar rules, is unlikely to demonstrate frequency effects. Essentially, given the effects in empirical studies, there is a need for language representation and processing models that account for frequency effects on compositional phrase processing.

The focus of the current study is on the processing of compositional multi-word sequences in language comprehension and production from the perspective of usage-based approaches to language acquisition. Researchers advocating these approaches, notably Nick Ellis and colleagues (e.g., Ellis, 2008a, 2008b, 2012, 2013; Ellis & Cadierno, 2009; Ellis, O'Donnell, & Römer, 2013; Ellis & Larsen–Freeman, 2009; Ellis

& Wulff, 2015; Robinson & Ellis, 2008), have proposed that L2 acquisition is also driven by the same general mechanisms that drive L1 acquisition. The current study is motivated by previous empirical evidence that native English speakers exhibited sensitivity to compositional word sequences, particularly sequences beyond the bigram (two word) level, in comprehension and production (e.g., Arnon & Priva, 2013; Arnon & Snider, 2010; Bannard & Matthew, 2008; Tremblay, Derwing, Libben, & Westbury, 2011). Another source of motivation was the growing number of studies demonstrating that L2 learners, particularly ESL learners, can retain memory of L2 word sequences and/ or were sensitive to the distributional properties of L2 input (e.g., Durrant & Schmitt, 2010; Ellis et al., 2008; Gyllstad & Wolter, 2016; Hernández, Costa, & Arnon, 2016; Sonbul, 2015; Wolter & Gyllstad, 2013). In the current study, a compositional multi-word sequence specifically refers to a word sequence consisting of at least three words (i.e., beyond the bigram level) and which has a meaning derivable from its parts (i.e., not proverbs, metaphors, or idioms). My objective is to investigate whether the claim put forth by usage-based researchers can be attested in both ESL learners' multi-word sequence comprehension and production.

CHAPTER 1: REVIEW OF THE LITERATURE

1.1 L1 acquisition from a usage-based perspective

A much-debated issue in the field of L1 acquisition has been the question of how humans acquire and process multi-word compositional sequences—sequences which have a meaning derivable from word parts. In the mainstream generative linguistics theory (Chomsky, 1995; Pinker, 1994, 1999), humans were born with a language acquisition device containing a set of rules governing human languages. This device is considered indispensible for children's L1 acquisition, as the input they receive should not be rich enough to enable them to acquire their L1 rapidly and uniformly as they do—that is, in this theory, there is a poverty of stimulus in L1 acquisition. From this perspective, there is also a separation of the lexicon and the grammar in a speaker's mental linguistic representation. The former constitutes an inventory of memorized non-compositional words (e.g., *cat, went*), bound morphemes (e.g., the past tense suffix *–ed*), and idioms (e.g., *for good*), while the latter consists of abstract morphosyntactic rules underlying the productive combination of lexical items into complex structures, including words, phrases, and sentences. For example, in English, an irregular past tense verb form (e.g., *went*) is stored as a memorized item in the mental lexicon, while the regular past tense verb form is generated based on the rule attaching the bound morpheme *–ed* to the end of a bare verb.

Generative linguists maintain that while the items in the mental lexicon have to be learned and memorized, children's acquisition of abstract L1 grammar rules result from exposure to input, which triggers the language acquisition device to set grammar rules specific to their L1. Moreover, learning lexical items and acquiring abstract grammar

rules depend on different cognitive abilities and even involve different parts of the brain (Ullman, 2001; Ullman et al., 2005). With regard to language processing, frequency effects, a psychological characteristic of memory, should be observed only from memorized items in the mental lexicon—namely, individual lexical items, bound morphemes, and idioms—but not with compositional phrases, which are computed real time based on abstract grammar rules during language processing (Prasada & Pinker, 1993; Prasada, Pinker, & Snyder, 1990; Ullman, 1999). Due to such a rigid distinction between the lexicon and the grammar and the two processing mechanisms, the processing model grounded in the generative linguistics theory has been commonly referred to the word-and-rule or the dual mechanism model.

Unlike the generative theory, a usage-based theory of language acquisition posits that L1 acquisition does not entail the innate language-specific acquisition device. Instead, a L1 is acquired on the basis of the interaction between language input and human domain-general cognitive processes, or cognitive processes that also operate in other areas of human activities (Abbot–Smith & Tomasello, 2006; Bybee, 2010; Goldberg, 1995, 2006; Tomasello, 2003, 2009). According to Bybee (2010), such cognitive processes include (1) categorization, or classification of a particular linguistic instance (e.g., *He kicked the ball*) into a particular type (e.g., a transitive sentence), (2) chunking, which creates sequential relations between co-occurring words, (3) rich memory, or storage of detailed information in previously-encountered linguistic input, (4) analogy, the creation of novel utterances based on similar existing linguistic patterns in the speaker's representation, and (5) form-meaning mappings. It is argued that language acquisition, structures, and processing that can be accounted for by these cognitive

processes have a psychological reality because these processes have been well-documented in non-linguistic human activities.

From a usage-based perspective, linguistic knowledge in a speakers' mind consists of constructions, or form-function mappings in a particular language. Thus, this perspective is compatible with cognitive linguistics (e.g., Croft & Cruse, 2004; Hoffmann & Trousdale, 2013; Langacker, 2008), including construction grammar (e.g., Goldberg, 2006), which view constructions as fundamental linguistic units. According to Goldberg (1995, 1999, 2003, 2006, 2013), the function of a construction involves its meaning. In English, for example, the transitive construction has the form of *Subject-Verb-Object* and its prototypical scene involves two participants, one acting on the other. The function of a construction sometimes also includes pragmatic or discourse information. For instance, one distinction between the active (e.g., *A car hit the armadillo.*) and the passive (e.g., *The armadillo was hit by a car*) is that the former topicalizes the actor (*a car*), while the latter topicalizes the undergoer of the action (*the armadillo*). Moreover and unlike the generative linguistics theory, usage-based views of language acquisition do not assume a complete division between the lexicon and the grammar in a speaker's linguistic representation. That is, the notion of constructions encompasses all types of form-function mappings. As Goldberg (2003, 2013) clearly pointed, constructions include (1) morphemes, (2) words, (3) idioms, (4) partially lexically filled linguistic patterns, in which some slots can be filled by various lexical items (e.g., *The __er the __er*), and (5) fully general linguistic patterns (e.g., the transitive pattern *Subject-Verb-Object*). Constructions therefore have varying levels of complexity and abstraction. Some examples of constructions are shown in Table 1.

Table 1. Examples constructions at varying levels of complexity and abstraction (Goldberg, 2003, 2013)

Type	Example
Morpheme	anti-__, pre-___, ___-ing
Word (partially filled)	avocado, anaconda, and
Idiom (filled)	going great guns
Idiom (partially filled)	jog ____ memory
Correlative construction	the __er the __er (e.g., *The more you think about it, the less you understand*)
Transitive construction	Subject-Verb-Object (e.g., *He kicked the ball*)
Ditransitive construction	Subject-Verb-Object1-Object2 (e.g., *She gave me a present*)

The creation and entrenchment of constructions is a speaker' mind emerges from generalizations of linguistic patterns made over more specific instances. That is, usage-based L1 acquisition is piecemeal in fashion. Ambridge and Lieven (2011) provided a clear example to illustrate this point. Initially, children are first exposed to specific instances of a construction (e.g., in case of the transitive construction, *John kissed Kate)*. Later, with more input and their memory storage of details about the construction—including the phonetic details, context of use, and associated meanings and inferences—children categorize and schematize across many instances of related utterances and form more abstract, yet still lexically-specific constructions (e.g., *KISSER kissed KISSEE*). Chunking is therefore important as it establishes sequential relations between co-occurring words, allows for the registration of a word sequence in memory, and thus makes the schematization possible. In addition, children's subsequent more exposure to input leads to their acquisition of more general, abstract adult-like constructions (e.g., *Subject-Verb-Object)*. Another assumption is that a speaker's utterances (e.g., *David kissed Liz*) are not always formed from the most abstract stored representation possible (e.g. *Subject-Verb-Object)*, but may be formed from any combinations of abstract and

more concrete relevant components (e.g., *KISSER kissed KISSEE, David kissed KISSEE).* In practice, the constructions a speaker stores as a mental representation depend on the input they have encountered. Moreover, the acquisition of a more general and abstract form does not mean that speakers cannot retain in their representation related, more lexically specific form—a simultaneous storage of more and less abstract forms is possible.

From usage-based views, construction acquisition is essentially an accumulation of statistical probabilities and abstraction of regularities out of construction occurrences in previously–encountered linguistic input. The acquisition is influenced by several psychological factors such as salience of the form and learner attention (e.g., Ellis & Larsen–Freeman, 2009; Gries & Ellis, 2015), and corpus linguists and psycholinguists have identified several types of statistical probabilities which should play a role in the acquisition, such as mutual information, an association strength between co-occurring words (Evert, 2008; Gries, 2010)[1], and delta P, which measures the probability of occurrence of a construction when a word is present minus when the word is absent (Ellis, 2006a; Ellis & Ferreira–Junior, 2009a; Gries, 2015). However, frequency is considered a key index of such probabilities, affecting the acquisition of constructions at all levels of complexity and abstraction (e.g., Ellis, 2002; Gries & Ellis, 2015). Drawing on Bybee and Thompson's (2000) work, Ellis (2002) distinguished between token frequency, how often a particular linguistic form occurs in the input, and type frequency, the number of different lexical items that can appear in a non-fixed slot in a construction

[1] An MI score is calculated by dividing the observed frequency of a word sequence in a specified span in a corpus by the corpus-based expected frequency and taking the logarithm to the base two of the result (Gries, 2010).

(e.g., *Subject-Verb-Object*). High token frequency promotes the entrenchment of words. Ellis (2002, 2005, 2012) additionally suggested that the cognitive ability of chunking allows strings of words (e.g., *David kissed Liz)* to be registered in human memory, and once the sequential relation is established, subsequent exposure to massive input leads to statistical fine-tuning, which makes the sequential relation reflect frequency of previous encounters. On the other hand, Ellis (2002) pointed out that type frequency determines the productivity of a more abstract construction. Speakers' exposure to a variety of lexical items in an unfixed slot in a construction (e.g., noun phrases with an actor role) leads them to form a general, more abstract category based on those items (e.g., *Subject*). Higher type frequency indicates more usage of an abstract construction. The representation of an abstract construction with a higher type frequency is more entrenched and is more accessible for further use with new lexical items. As Ambridge and Lieven (2011) pointed out, speakers can simultaneously retain the representation of both a more general and abstract construction and more lexically specific related forms. Thus, in light of these theoretical proposals, speakers should exhibit sensitivity to frequency of compositional multi-word sequences they have encountered, which have left traces in their rich memory, even though they have acquired more abstract related forms.

While researchers informed by usage-based approaches hold similar views regarding the general mechanism of language acquisition, Arnon and Snider (2010) pointed out that these researchers have two different assumptions about the nature of compositional multi-word sequence representation. One assumption—grounded in the work of researchers such as Goldberg (2006) and Wray (2002)—is that "phrases that are of sufficient frequency can attain independent representation as a way of making

processing more efficient" (Arnon & Snider, 2010, p.69). While there is not yet a clear consensus as to the minimum frequency level (i.e., sufficient frequency), in this view, there is a qualitative difference between highly frequent phrases (i.e., stored as a whole) and less frequent phrases (i.e., generated or analyzed by the language grammar), and the first type of phrases are processed faster by the second. By contrast, a different assumption—informed by work of researchers such as Bybee and colleagues (Bybee, 2006, 2010; Bybee & Hopper, 2001)—is that there is no such a qualitative difference. Speakers retain information about all compositional phrases they have been exposed to. More frequent phrases are more entrenched in speakers' representation; the difference between a more frequent phrase and a less frequent phrase is quantitative, resulting from different frequencies of previous phrase encounters. Therefore, phrases do not need to be highly frequent to be processed faster; relatively more frequent phrases should be processed faster than less frequent phrases regardless of the frequency range.

Despite the different views, a common prediction from a usage-based perspective is that frequency effects can be observed from compositional multi-word sequence processing. This prediction thus differs from the prediction in the generative linguistics theory, which is that frequency effects should be observed only with items stored in the mental lexicon (i.e., bound morphemes, individual words, and idioms), and not with compositional word sequences, which are considered to be computed real-time based on abstract grammar rules (e.g., Prasada et al., 1990; Prasada & Pinker, 1993; Ullman, 1999). In the current study, by recognizing that usage-based approaches, which have recently amassed L2 research attention (e.g., Ellis & Larsen–Freeman, 2009; Eskildsen, 2012; McDonough, & Nekrasova–Becker, 2012; McDonough, & Trofimovich, 2013;

Ortega, 2013; Ortega, Tyler, Park, & Uno, 2016; Robinson & Ellis, 2008; Römer, O'Donnell, & Ellis, 2014; Tyler, 2010; Wulff, Ellis, Römer, Bardovi–Harlig, & Leblanc, 2009), have both theoretical strengths and remaining issues to explain (Ibbotson, 2013), I tested whether ESL learners exhibit sensitivity to compositional multi-word sequence frequency in both comprehension and production. In doing so, I therefore sought evidence for the prediction made by usage-based researchers in L2 acquisition. Moreover, given the possible differences in the input and nature of L1 and L2 learning (Ellis & Larporte, 1997; Bley–Vroman, 2009; DeKeyser, 2000; Muñoz, 2008), if frequency effects are psychologically real in L2 learners, the implication is that such sensitivity should be accounted for by L2 representation and processing models. It should also be pointed out that, as discussed, in addition to frequency, there are other statistical measures, such as mutual information, which reflect the statistical properties of previously encountered multi-word sequences. While an investigation into sensitivity to these measures is also interesting from a usage-based perspective, frequency is the only focus of the current study.

1.2 Previous work on frequency effects on compositional word sequence processing in native English speakers

Much previous research has demonstrated frequency effects at varying degrees on compositional phrases in the receptive language processing of native English speakers. The existing evidence is from various types of tasks, including phrasal decision tasks, in which participants judged whether the stimuli were possible English word sequences (Gyllstad & Wolter, 2016; Wolter & Gyllstad, 2013), self-paced reading tasks (Reali & Christiansen, 2007; Tremblay et al., 2011), word-monitoring tasks (Sosa & Macfarlane,

2002), and reading with concurrent eye-movement registration (Siyanova–Chanturia, Conklin, & van Heuven, 2011; Sonbul, 2015). While the results from these studies seemed compatible with usage-based approaches, Arnon and Snider (2010) identified two limitations of the existing empirical work in this area. First, the focus of most studies was only on two-word compositional phrase processing (e.g., Sonbul, 2015, Sosa & Macfarlane, 2002), but stronger support for usage-based approaches would be from frequency effects on longer compositional phrases. Second, in previous studies on longer sequences (e.g., Siyanova–Chanturia et al., 2011; Tremblay et al., 2011), the frequency of subparts of stimuli was not strictly controlled for. This therefore may cast doubt on whether the processing differences between higher and lower frequency sequences could be fully attributable to the frequency of the whole sequences. To address these limitations, Arnon and Snider (2010) used a phrasal decision task to investigate frequency effects on compositional four-words-sequence recognition. Their stimuli were pairs of phrases, each consisting of two phrases with the same words except the last (e.g. *don't have to worry* vs. *don't have to wait*, the former being more frequent and the latter less frequent). Their analysis strictly controlled for the frequency of the subparts of the stimuli. The results demonstrated that reaction times to more frequent phrases were significantly shorter than reaction times to less-frequent ones. This was the first study that strictly controlled for substring frequency and supported the psychological reality of frequency effects on compositional multi-word sequence recognition beyond the bigram level in adult native English speakers. A very recent study by Hernández et al. (2016), which used the same type of task and controlled for substring frequency, also documented similar effects. These two studies thus lend stronger support for the

prediction in usage-based views of language acquisition.

Some other studies investigated frequency effects on composition phrase production in native English speakers. Such studies have demonstrated the effects on phonetic reductions (Aylett & Turk, 2004; Bybee & Scheibman, 1999; Jurafsky et al., 2000) and voice onset time (VOT[2])—that is, when native English speakers saw word sequences, they started to produce more frequent word sequences faster than when they saw less frequent ones (Jannsen & Barber, 2012). Other studies looked at frequency effects on elicited compositional multi-word sequence production beyond the bigram level in an experimental setting, but these studies have yielded mixed results. In one study, Bannard and Matthew (2008) used an oral repetition task with 38 children aged 2-3 years old. The stimuli were pairs of four-word phrases differing only in the last word (*when we go out* vs *when we go in*), constructed from the 1.72-million-word Max Planck Child Language Corpus, a corpus of speech directed to a child when he was between the ages of 2 to 5. In each stimuli pair, one phrase (e.g., *when we go out*) was more frequent than the other (e.g., *when we go in*). During the experiment, each participant sat in front of a computer screen on which a picture of a tree with stars in the branches was presented. The participant was instructed to (1) listen to what the computer said and (2) "say the same thing" (Bannard & Matthew, 2008, p. 44) to get a sticker to cover each of the stars. An experimenter clicked a mouse to play one target phrase at a time. The outcome measures were production accuracy and production durations of the first three words, which were identical in each target pair (*e.g., <u>when we go</u> in vs <u>when we go</u> out).* If the children did not say a phrase, the experimenter also prompted them to respond,

[2] In the current paper, based on Ellis et al.'s (2008) definition, VOT is the duration between the onset of stimuli and the beginning of a participant's oral response.

such as by saying *Can you say that?* However, if the children still did not say the phrase, the experimenter skipped to the next audio clip. The results revealed that the children were significantly more likely to say higher frequency phrases correctly. Also, whole phrase frequency was a significant predictor of production durations when stimuli substring frequencies were controlled for.

With regard to research with adult native English speakers, Tremblay and Tucker (2011) used an elicited production task, in which participants saw 432 compositional four-word phrases on a computer screen (e.g., *I don't really know*), one phrase at a time, and said the phrases as fast as they could after the phrases appeared on the screen. The results revealed that phrase frequency did not significantly predict the participants' speech durations. However, one methodological issue that could have affected the results was how Tremblay and Tucker entered some control variables into their regression analyses. For example, the researchers included (1) the interaction between frequency of the first word and the frequency of the third word in the target four-word phrases, and (2) the interaction between whole phrase frequency and the frequency of the first two words in the target phrases. These interactions were significant in the analysis, but what these interactions mean were not clearly explained. In another study, Ellis et al. (2008) used a similar elicitation task. Their phrases were compositional three- to five-word academic sequences (e.g., *see for example, it has been shown, it should be noted that*) sampled from the Michigan Corpus of Academic Spoken English, selected academic written and spoken files from the British National Corpus (BNC; Leech, 1992), and Hyland's (2004) corpus of academic research articles. The results also suggested that adult native speakers did not exhibit frequency effects. However, in this study, there was a lack of

substring frequency control. For example, production durations of higher frequency phrases such as *it can be seen that* were compared against production durations of less frequent phrases such as *as in the case of*. It was unclear, therefore, how the frequency of *it can be* and *as in the* affected the production durations. Given the results from these studies with native English speakers, one possibility was that frequency effects on sequences beyond the bigram level may be present only in children (Bannard & Matthew, 2008). The second possibility was that the incongruent findings stemmed from methodological issues, namely, the inclusion of control variables that were difficult to be interpreted (Tremblay & Tucker, 2011) or the lack of substring frequency control (Ellis et al., 2008).

To investigate this latter possibility, in one of their experiments, Arnon and Priva (2013) used a similar elicited oral production task, in which the stimuli were a subset of the target pairs from Arnon and Snider (2010). In each pair, one phrase was classified as a high frequency phrase (e.g., *don't have to worry*) and the other a low frequency phrase (e.g., *don't have to wait*). Participants saw one four-word phrase at a time on a computer screen and were instructed to say the phrase as fast as they could once the phrase disappeared from the screen. The outcome measure was production durations of the first three words in the target phrases (e.g., *don't have to*). Arnon and Priva (2013) found shorter phonetic durations for three-word sequences embedded in higher frequency phrases. Like Bannard and Matthew (2008), Arnon and Priva (2013) controlled for relevant substring frequencies in their analysis. However, Arnon and Priva's (2013) elicited production task differed from Bannard and Matthew's (2008) task in two ways. First, Bannard and Matthew (2008) merely instructed the children to repeat after the

phrase after it was played, and the researchers tried to elicit the children's response once if the children did not repeat after the phrase. Arnon and Priva (2013), on the other hand, instructed the participants, who were adults, to say each phrase as fast as they could after the phrase disappeared from the screen. Second, Bannard and Matthew (2008) used a within-subject design, in which all the children were exposed to all the target phrases. Therefore, production of the higher and lower frequency phrases were from the same participants, and as a result individual variability in speech production speed was controlled for. By contrast, Arnon and Priva (2013) used a between-subject design to avoid a repetition effect resulting from a participant's reproduction of the same trigrams from a target pair (e.g., *don't have to worry* and *don't have to wait*). Consequently, each participant read only one variant from each target pair (i.e., one participant's production of *don't have to* in *don't have to worry* was compared against another participant's production of this trigram in *don't have to wait*). To control for individual variability in production speed, Arnon and Priva (2013) entered the average production durations of each participant (across all target stimuli) in their regression model as a predictor of production durations to account for individual variability. Despite these differences, the results from the elicited production tasks in the two studies appeared to support the usage-based prediction about frequency effects.

Finally, two studies demonstrated frequency effects on corpus-derived spontaneous compositional multi-word sequence production in adult native English speakers. The first study was one of the studies conducted by Arnon and Priva (2013).[3]

[3] This study included an elicited production experiment and a separate spontaneous speech production experiment. These two experiments were based on different participants.

Their target phrases were three-word sequences of two types, *Subject–Auxiliary–Verb* (e.g., *everybody was trying*) and *Verb–Determiner–Noun* (e.g., *saw the boy*), selected from the Switchboard Corpus of Spoken American English (Godfrey, Holliman, & McDaniel, 1992). The researchers strictly controlled for substring frequency, obtained from the Switchboard and the Fisher corpora (Cieri, Miller, & Walker, 2004) combined. The results revealed that phrase frequency significantly predicted production durations of the target phrases. A similar result was obtained by Arnon and Priva (2014) from their analysis of the Buckeye Corpus of Conversational Speech (Pitt et al., 2007). Their target phrases were three-word sequences in which the word in the middle is a noun.[4] Due to a relatively small size of the corpus (around 300,000 words), the researchers again controlled for substring frequency in their analysis by deriving substring frequency from the Switchboard and the Fisher corpora. In the spontaneous speech analyses in these two studies, one issue that needed to be clarified so that the observed frequency effects can be cited as strong evidence for usage-based approaches is whether all the analyzed phrases were compositional phrases (e.g., did not include a sequence embedded in an idiom). Overall, however, the results seemed to corroborate the ontological status of frequency effects in adult native speakers' language production of compositional English multi-word sequences.

To summarize, the prediction in usage-based approaches about compositional phrase frequency effects has been borne out at various degrees in previous empirical research on L1 comprehension and production. The supporting evidence seems to

[4] According to Arnon and Priva (2014), this type of phrase was selected because nouns constituted the most diverse word class (in terms of the number of types) in this corpus. However, no specific example of the selected phrases was provided.

challenge a language processing model predicting frequency effects only on the processing of individual lexical items, bound morphemes, and idioms. Usage-based approaches have also motivated a great deal of L2 acquisition research, which I describe next.

1.3 Frequency in L2 acquisition

The amount of L2 learning, teaching, and psycholinguistic research grounded in usage-based approaches has surged in recent years. This followed from the theoretical argument that, as in L1 acquisition, an L2 may be acquired on the basis of language input and domain general cognitive processes (Bybee, 2008; Ellis, 1996, 2002, 2003, 2006a, 2006b, 2008a, 2008b, 2011, 2012, 2013; Ellis & Cadierno, 2009; Ellis et al., 2013; Ellis & Larsen–Freeman, 2009; Ellis & Wulff, 2015; Goldberg & Casenhiser, 2008; Robinson & Ellis, 2008). Ellis (2002) stated clearly that, "[T]he L1 acquisition sequence … could serve well as a reasonable default in guiding the investigation of the ways in which exemplars and their type and token frequencies determine the [L2] acquisition of structure" (p.170). That is, as in L1 acquisition, usage-based L2 acquisition is also driven by accumulation of statistical probabilities of previously-encountered L2 input. Ortega (2013), in her relatively recent critical review of L2 acquisition research, identified the growth of usage-based L2 research and its connection to other theoretical frameworks which had previously informed only L1 research—such as connectionism, construction grammar, cognitive linguistics, and complex adaptive system—as an important L2 research trend in the 21th century. The growth and the connection, Ortega (2013) maintained, had a potential to enhance our understanding of human cognition and

language science in general. I therefore believe that an investigation of frequency effects on L2 composition multi-word sequence processing is not only timely but also necessary because the effects constitute an important piece of psychological evidence for usage-based L2 acquisition.

Based on usage-based views of language and acquisition, the implication for the current study, which is in the context of ESL, is that frequency effects should be observed in ESL compositional multi-word comprehension and production. However, the L2 acquisition literature has documented numerous possible differences between L1 and L2 acquisition. Thus, even if usage-based researchers' claim about the general language mechanisms is true, an empirical inquiry into the effects in adult ESL learners may illuminate the differences between L1 and L2 learning if the effects are found to be different from those in adult native English speakers. Native – non-native differences can be expected for several reasons. First, the amount, quality, and structure of input that native English speakers and adult L2 learners receive may differ. Ellis and Laporte (1997), for example, observed that English input in formal English classrooms differ from child-directed L1 input in several ways, including the amount of input and the nature of the interactions (e.g., naturalistic L1 exposure vs explicit instruction in L2 classrooms). Likewise, drawing on corpus data, Littlemore (2009) reported that L1 and L2 construction acquisition may differ because L1 child-directed speech and English input that L2 learners receive through interactions with adult native speakers may contain dissimilar specific instantiations of a particular linguistic construction. In addition, according to Muñoz (2008), compared to input that native speakers receive, input in formal ESL classrooms has less quantity, and such ESL input deprives ESL learners of

optimal L2 learning conditions because neurolinguistic research has suggested that input intensity is indispensible for the development of neural representation of multiple languages. In light of these observations, it is therefore possible that frequency will affect compositional multi-word processing in native English speakers and ESL learners differently.

A second possible factor that may lead to different frequency effects between the two groups pertains to the possible differences between the nature of L1 and L2 acquisition. While sensitivity to statistical information in previously-encountered L2 input is generally considered to result from implicit learning mechanisms (e.g., Ellis, 2002, 2013; Ellis & Larsen–Freeman, 2009), unlike child L1 learners, older L2 learners may be less apt at acquiring linguistic patterns implicitly and rely more on explicit learning (Bley–Vroman, 2009; DeKeyser, 2000). According to some researchers who empirically investigated initial L2 verb argument construction acquisition from a usage-based perspective, this possible L1-L2 acquisition difference may have been responsible for their research findings (McDonough & Nekrasova–Becker, 2014; McDonough & Trofimovich, 2013; Nakamura, 2012; Year & Gordon, 2009). As discussed, in usage-based views, a speaker's acquisition of an abstract verb argument construction results from schematization across specific instances of that construction in previously encountered input. One further observation (Goldberg, 1999; Goldberg, Casenhiser, & Sethuraman, 2004) is that, in L1 child-direct speech, the distribution of verbs in a verb argument construction generally adheres to the Zipf's (1935) law, whereby frequency of the verbs in a construction (e.g., the English ditransitive construction, *Subject-Verb-Object1-Object2*) declines as a power function of their frequency rank in that particular

construction. That is, there is a strong tendency for one verb (e.g., *give*) to occupy a disproportionally large share in the distribution, and subsequent verbs (e.g., *pass, sell, throw*) in the frequency rank have rapidly declined frequency. The highly frequent verb is typically the verb that can be applied to various situations and convey the meaning closely associated with the prototypical meaning of the construction (e.g., *possession transfer*). Interestingly, this input pattern is also the pattern in children's speech production. These corpus-based observations have motivated a hypothesis that verb argument construction acquisition is facilitated by low-variability input, which contains a few instances of possible verbs in that construction, particularly low-variability input with a skewed pattern of distribution, in which one verb appears disproportionally more frequent that the other possible verbs. The reason is because the skewed pattern facilitates speakers' detection of the underlying, more abstract verb argument construction (Goldberg, 1999; Goldberg et al., 2004).

This hypothesis has been attested in experimental L1 studies on initial verb argument construction acquisition. That is, when native English speakers learned a novel verb argument construction containing novel verbs, skewed input led to a better learning outcome than balanced input, in which different verbs appeared with an equal frequency in a target construction (Casenhiser & Goldberg, 2005; Goldberg et al., 2004). The skewed input effect also seemed greater when the disproportionally high frequency verb appeared consecutively first before other verbs in a target construction appeared (as opposed to when the high frequency verb was randomly interspersed throughout input sentences) (Goldberg, Casenhiser, & White, 2007). These corpus and experimental findings have motivated numerous usage-based L2 studies in which adult L2 learners

were exposed to L2 input in training sessions and the learners' initial construction acquisition was subsequently assessed with tasks such as a sentence acceptability judgment task or a listening comprehension task. However, unlike the L1 studies (e.g., Goldberg et al., 2004), such L2 studies have repeatedly documented no significant difference in the effectiveness between skewed and balanced input (Fulga & McDonough, 2016; Nakamura, 2012) or reported the superiority of balanced input (McDonough & Nekrasova–Becker, 2014; McDonough & Trofimovich, 2013; Year & Gordon, 2009). This has prompted a speculation that one cause for the incongruent L1-L2 findings may be adult L2 learners' use of more explicit learning strategies, compared to native speakers' learning of input-based linguistic patterns at a more implicit level (McDonough & Nekrasova–Becker, 2014; McDonough & Trofimovich, 2013; Nakamura, 2012; Year & Gordon, 2009).

However, some caution should be exercised when the results from such L2 studies are interpreted. In addition to the possibility that L2 learners are less apt at learning an L2 implicitly, other factors may have contributed to the incongruence between the L1 and L2 findings. For example, in previous L1 studies on skewed input by Casenhiser and Goldberg (2005), participants, who were native English speakers, learned a novel English construction, *the appearance construction*, which contained novel verbs—such as *moopo* and *feg*—and had a novel form of *NounPhrase1–NounPhrase2–Verb,* as in *A rabbit the hat moopoed* (A rabbit appears on a hat), and *The sun the sky fegoed* (The sun rises into a sky). On the other hand, in an L2 study by McDonough and Nekrasova–Becker (2014), which reported no skewed input effect, the researchers acknowledged that the Thai English learners in their study may have had prior experience

with the target English structure, the ditransitive construction, before participating in the experiment. The absence of the skewed input effect was also documented in McDonough and Trofimovich's (2013) study, in which Thai university students learned the transitive construction in Esperanto, characterized by a suffix added to an object noun.[5] As McDonough and Trofimovich (2013) conceded, in the previous L1 studies demonstrating the skewed input effect (e.g., Casenhiser & Goldberg, 2005), participants learned a construction associated with a verb, and the disproportionally frequent verb in the skewed input condition conveyed the most prototypical meaning of that construction. By contrast, the Esperanto transitive construction is associated with a noun suffix and, in McDonough and Trofimovich's (2013) study, the Esparanto noun presented with high frequency in their skewed input condition did not convey the prototypical meaning of an object (an inanimate noun receiving an action from a human agent). In another L2 study, Year and Gordon (2009) speculated that one possible reason for the absence of skewed input effect might have been the presentation of sentences in their training sessions. That is, while the researchers' only target construction was the English ditransitive, which their participants had never learned prior to the experiment, the researchers presented a target verb (e.g., *give*) in both the ditransitive construction (e.g., *Subject-Verb-Object1-Object2,* as in *Peter gave Karen a book*) and the alternate prepositional dative construction (e.g., *Subject-Verb-Object1-to-Object 2*, as in *Peter gave a book to Karen*) in the training sessions. Year and Gordon (2009) thus conceded that this might have created some noise that affected their results. Due to these various differences between the L1 and L2 studies on skewed input, it is difficult to draw a solid conclusion that the

[5] Strictly speaking, this is also third language acquisition research.

absence of the skewed input effect in L2 learners demonstrated a contrast in the nature of initial L1–L2 construction acquisition, let alone the conclusion that adult L2 learners cannot implicitly accumulate statistical information in previously-encountered input after a great deal of L2 exposure.

Furthermore, one specific possible difference between L1– L2 acquisition concerns L2 learners' memory retention of L2 word co-occurrences. Contrary to Ellis (1996, 2002), Wray (2002) hypothesized that, unlike L1 learners, L2 learners may not retain memory about L2 word co-occurrences but instead break phrases into individual words. This results from their lack of necessity to memorize and use frequently occurring L2 word sequences, their L2 education, which typically focuses on forms and individual words, and their mature cognitive development and L1 literacy, which prompt them to break down lexical sequences into words. Moreover, although L2 learners may intentionally memorize frequent L2 word sequences, the resulting knowledge may not be attuned to the statistical properties of the L2 input. However and as I will discuss in a subsequent section, empirical research demonstrating frequency effects on ESL compositional word sequence comprehension (e.g., Sonbul, 2015; Wolter & Gyllstad, 2013) seem to provide counterevidence to Wray's (2002) claim.

Despite the possible differences between L1 and L2 acquisition and unlike Wray (2002), several usage-based researchers (e.g., Ellis, 1996, 2003, 2011, 2012; Hernández et al., 2016) have suggested that frequency effects may be observed in ESL learners. The reason is because, as in L1 acquisition, the human cognitive ability of chunking allows L2 word sequences to be registered in L2 learners' memory, thereby creating sequential relations between words in the sequences. With implicit processing of registered word

strings in massive subsequent L2 input, the strength of the relations will reflect the frequency of previous encounters. Describing the emergence of L1 and L2 linguistic representation in a speaker's mind, Ellis (2012, p. 25) stated this point clearly:

> "Language users (both L1 and L2) are sensitive to the sequential statistics of these dependencies, large and small…The results encourage an emergentist view whereby all linguistic material is represented and processed in a similar fashion, where learners are sensitive to the frequencies of occurrence of constructions and their transitional probabilities, and hence where they have learned these statistics from usage, tallying them implicitly during each processing episode."

In light of this claim, in the case of the adult ESL learners who are university students in the US, the target participants in the current study, these learners may demonstrate frequency effects because they have lived in the English speaking environment for a certain period of time.[6] Moreover, considering the fact these learners had to develop a sufficient level of English proficiency for US university admissions, it is reasonable to assume that, before coming to the US, these learners had been exposed to English from various sources (e.g., TV programs, news reports, movies). Possibly, English phrases that are more frequent in an English environment are also more frequent in those sources, although this is a possibility that needs further empirical support. In addition, these adult ESL learners typically had received a great deal of English instruction in formal classrooms in their home country, and such instruction usually involved explicit instruction (e.g., Ellis & Laporte, 1997). Ellis (2005, 2011, 2012) posited that explicit instruction can also play a role in usage-based L2 acquisition. That is, the acquisition may be facilitated by instruction or practice that draws ESL learners'

[6] It is difficult to point out the exact amount of time because I am aware of no literature specifically discussing how long it is for frequency effects to emerge in adult L2 learners.

attention to word sequences because such a pedagogical intervention helps register a novel L2 word sequence in a learner's memory before the subsequent fine-tuning through encounters of the sequence in more L2 input. Whether word sequences which are more frequent in an English speaking environment are also more likely to be explicitly taught in formal English classrooms also requires empirical evidence. However, based on these proposals and possibilities, the effects of frequencies derived from native speaker corpora may be observed in adult ESL learners who are proficient enough to study in an English environment and who have lived in an English speaking environment for a certain period of time. Indeed, such effects seemed to be observed in several previous L2 studies discussed in the next section.

Perhaps it should also be pointed out that one issue that has not been specifically discussed in great detail from a usage-based perspective is the potential role of written input in L2 acquisition—that is, whether and to what extent the representation of L2 compositional word sequences is influenced by frequency of encounters in previous L2 reading. This area of discussion is particularly relevant in the domain of L2 acquisition research because arguably a great deal of L2 input is in the written form (e.g., L2 or academic textbooks, magazines, news articles). To date, some studies on ESL collocation learning have suggested the possibility that ESL learners can retain memory about L2 word co-occurrences encountered in previous reading. Durrant and Schmitt (2010), for example, specifically tested the contradictory predictions by Ellis (1996, 2003) and Wray (2002) regarding adult ESL learners' memory retention of word co-occurrences. In a training session, participants read target adjective-noun collocations (e.g., *suitable wine*) and distractors on a computer screen, and in a subsequent testing

session, the participants completed a naming task, in which they were presented with adjectives from the target collocations (e.g., *suitable*) and the first two letters of the following noun (e.g., *suitable WI_ _*). In the test, the participants were informed that the noun appeared in the training session and they were instructed to say the noun if they knew it. The results revealed that, in the testing session, the participants were significantly more likely to remember a noun (e.g., *wine*) if the noun appeared with the adjective prime (e.g., *suitable*) in the training session. Moreover, the memory retention increased as a function of the number of times the adjective-noun collocations appeared in the training session. Similarly, Webb, Newton, and Chang (2013) investigated adult English learners' incidental learning of 18 verb-noun collocations embedded in a graded reader. Four versions of the graded reader were created, each differing in the number of times each target collocation appeared. The results indicated a positive and significant correlation between collocation learning gains and the number of collocation encounters in the texts. These two studies thus seemed to suggest that word sequence encounters in reading can leave some traces in adult ESL learners' memory and that the frequency of exposure influences the strength of association between component words in compositional sequences. Therefore, the results, although based only on immediate posttests, appeared in line with Ellis's (1996, 2003) prediction. To date, however, a direct discussion about the role of written texts when compared to the role of speech input in usage-based L2 acquisition seems scarce. This is perhaps not surprising because L2 acquisition theories—whether usage-based or generative—originate from child L1 acquisition theories (e.g., Chomsky, 1995; Tomasello, 2003, 2009), which attribute language acquisition to exposure to spoken input and subsequent implicit acquisition.

Interestingly, despite this limited discussion, in several ESL processing studies on frequency effects, to be discussed in the following section, the stimuli or frequency data were derived from both spoken and written corpora, such as a written component of the BNC or the Corpus of Contemporary American English (COCA; Davies, 2013). This seems to suggest that frequency of encounters during reading may influence L2 representation from a usage-based perspective, although at this point encounters in spoken input seems most relevant given the theoretical influence from child L1 acquisition theories.

To summarize, in spite of several possible differences between L1 and L2 acquisition, in usage-based approaches, L2 learners may exhibit frequency effects when processing L2 compositional multi-word sequences. This prediction has spurred an expansion in empirical studies on frequency effects in ESL learners, which are discussed next.

1.4 Previous work on frequency effects on ESL word sequence comprehension

There has been quite a great deal of empirical research showing frequency effects on ESL single word comprehension (Diependaele, Lemhöfer, & Brysbaert, 2012; Duyck, Vanderelst, Desmet, & Hartsuiker, 2008; Gollan, Montoya, Cera, & Sandoval, 2008; Whitford & Titone, 2012). While relatively more limited, empirical investigation into receptive processing of frequently co-occurring ESL word sequences has also grown in recent years (for reviews see Ellis, 2012; Siyanova–Chanturia & Martinez, 2014). Referred to in many ways, these sequences (Wray, 2002) can be broadly defined as those that "have become conventionalized in a given language as attested by native-speaker

judgment and/or corpus data" (Boers & Lindstromberg, 2012, p. 83). Many of these studies, however, did not specifically investigate the contradictory predictions about frequency effects on compositional phrases made by usage-based and generative linguists. That is, some existing research investigated the processing advantage of idioms (e.g., *hit his head on the nail)* over control phrases (e.g., *hit the nail on the head*), as measured by shorter reading time in a self-pace reading task (Conklin & Schmitt, 2008; Schmitt & Underwood, 2004) or by eye movements (Siyanova–Chanturia, Conklin, & Schmitt, 2011; Underwood, Schmitt, & Galpin, 2004). Another type of research focused on whether ESL learners processed literal meanings of English idioms faster than figurative meanings (Cieślicka, 2006). Yet the other type of research sought to investigate the processing advantage of sequences with semantic coherence or rhetorical functions usually taught as fixed chunks to ESL learners (e.g., *to begin with*) over control phrases (e.g., *to dance with*) (Jiang & Nekrasova, 2007). Frequency effects on compositional multi-word phrases were not the focus in these studies.

Other studies, listed in Table 2 based on stimuli type, investigated frequency effects on ESL phrase comprehension. First, a study by Kim and Kim (2012) was unique in that it investigated the processing of phrasal verbs consisting of a verb plus the particle *out* (e.g., *work out, bail out, wear out*) by native English speakers and advanced ESL learners in a self-paced reading task. The stimuli, derived from COCA, were divided in four frequency categories: (1) low, (2) mid-low, (3) mid-high, and (4) high. Each phrasal verb was embedded in a context sentence, which was presented in a segment-by-segment self-paced moving window format. Each target phrasal verb was always presented as a single segment. Based on reading times, the native English speakers processed phrasal

Table 2. Previous studies investigating frequency effects on ESL compositional phrase comprehension

Study	Type of phrases	Participants	Participants' L1(s)
Kim and Kim (2012)	Two-word phrasal verbs (e.g., *find out*)	• Students at a US university (*N*=14) • Relatively advanced proficiency, measured by internet-based TOEFL scores (*M* = 107; no *SD* provided)	Chinese, Korean, Japanese
Wolter and Gyllstad (2013)	Adjective-noun collocations	• University students in Sweden (*N*=25) • Advanced English proficiency, measured by Eurocentres Vocabulary Size Test (Meara & Jones, 1990)	Swedish
Sonbul (2015)	Adjective-noun collocations	• Students at a UK university (*N*=30) • Relatively proficient in English; minimum paper-based TOEFL score of 550 or IELTS score of 6.0	15 different L1s (e.g., Arabic, Chinese, German, Thai)
Gyllstad and Wolter (2016)	Verb-noun sequences	• University students in Sweden (*N*=27) • Advanced English proficiency, measured by Y_Lex Test of Vocabulary Size (Meara, 2005)	Swedish
Ellis et al. (2008)	Three- to five-word phrases	• Students enrolled in English for academic purpose classes at a US university (*N*=11) • Relatively proficient in English, but proficiency measures not specified	Chinese, Thai, Korean, and Spanish
Siyanova–Chanturia, Conklin, and van Heuven (2011)	Three-word phrases	• Students at a UK university (*N*=28) • Relatively proficient in English, but proficiency measures not specified	Various unspecified L1s
Valsecchi et al. (2013)	Four-word phrases	• University students in Germany (*N*=15) • English proficiency not clearly identified but presumably advanced learners excluded	German
Hernández et al. (2016)	Four-word phrases	• Students at a US university (*N*=27) and students majoring in translation and interpretation at a Spanish university (*N*=25) • Either upper intermediate or lower advanced English proficiency, measured by Lexical Test for Advanced Learners of English (Lemhöfer & Broersma, 2012)	12 different L1s in the first group (e.g., Spanish, Chinese, Italian) and Spanish or Catalan in the second group

verbs in the low frequency category significantly more slowly than phrasal verbs in each of the other three categories, and the ESL learners processed the phrasal verbs in the low frequency category significantly more slowly than phrasal verbs in the high frequency category. No other significant differences between frequency categories were observed in any of the two participant groups. Consequently, Kim and Kim (2012) argued that the native speakers stored and processed only phrasal verbs in the mid-low, mid-high, and high frequency categories holistically (i.e., as chunks), so the reading times for stimuli in each of these categories were significantly shorter than the readings times for stimuli in the low-frequency category. Their additional argument was that in the ESL learners, only phrasal verbs in the high frequency category were stored as chunks; these phrasal verbs "are most widely used and encountered in [the ESL learners'] daily life" (p. 838). Kim and Kim (2012) also concluded that their results were compatible with the dual mechanism model of language processing (Pinker, 1999; Pinker & Ullman, 2002), extending frequency effects on single words, bound morphemes, and idioms to phrasal verbs. This conclusion seemed to suggest that a phrasal verb is stored as a memorized chunk in the mental lexicon. However, Kim and Kim's (2012) conclusions need further clarification. First, in the native speaker group, if phrasal verbs in the mid-low, mid-high, and high frequency categories are indeed stored holistically, it was unclear why there was no significant difference in the reading times between any two of these three categories (i.e., mid-low vs. mid-high, mid-high vs. high) because, in the dual mechanism model, items stored holistically in the mental lexicon are predicted to demonstrate frequency effects (e.g., Prasada et al., 1990; Prasada & Pinker, 1993; Ullman, 1999). Second, Kim and Kim's (2012) claim that some but not all phrasal verbs are stored holistically is not

yet an established analysis in the generative approach, as there are still several proposals as to how phrasal verbs are stored and represented (den Dikken, 1995; McIntyre, 2001).[7] Furthermore, other factors besides phrase frequency may have affected Kim and Kim's (2012) results, including the influence of context sentences, the ESL participants' different L1 backgrounds, and whether the learners knew all the component words in the stimuli. For example, the verb *wander* (COCA frequency = 3,710 occurrences) in the target phrasal verb *wander out* had a much lower frequency than the verb *clean* (COCA frequency = 41,978 occurrences) in *clean out.* If the ESL learners did not know a low frequency word such as *wander*, the lack of familiarity with component words, not just phrasal verb frequency, could have affected the results.

Three other studies (Gyllstad & Wolter, 2016; Sonbul, 2015; Wolter & Gyllstad, 2013) investigated frequency effects on the processing of two-word ESL collocations, and these studies have lent varying degree of support for usage-based approaches. First, Wolter and Gyllstad (2013) used a phrasal acceptability task to investigate frequency effects on adjective-noun collocation (e.g., *middle class, commercial break*) comprehension in advanced Swedish learners of English. The results showed the learners' accuracy increased and reaction times decreased as a function of COCA frequency. Importantly, the frequency effects were observed after the frequency of the subparts of the target collocations—namely, the adjectives and the nouns—were controlled for, suggesting that the effects were driven by whole phrase frequency. In another study, Sonbul (2015) used a reading task with concurrent eye-movement registration to investigate comprehension of the same type of collocations. Derived from

[7] I thank Patti Spinner for her suggestion about this.

the BNC, the stimuli included thirty sets of higher frequency collocations (e.g., *fatal mistake*), lower frequency collocations (e.g., *awful mistake*), and unidiomatic sequences (e.g., *extreme mistake*). The stimuli in each set shared the same noun (e.g., *mistake*) and contained semantically related adjectives (e.g., *fatal, awful, extreme*). The sequences from each set were embedded in the same context sentence (e.g., *The engineer made one fatal/ awful/ extreme mistake which weakened the bridge*), and the resulting three sentences from each set were distributed over across three counterbalanced experiment blocks. The results revealed that adult native English speakers and ESL learners demonstrated sensitivity to collocation frequency as measured by first pass reading time; however, the effects were not observed based on total reading time and fixation counts, suggesting that collocation frequency affected only initial reading. Sonbul (2015) speculated that the absence of frequency effects based on these two measures may have resulted from the fact that adjective-noun collocations are not completely fixed and that, after initial collocation encounters, "language users might be more tolerant of alterations in their structure" (p.13). There were also some methodological issues that may affected the results. In the case of the ESL learners, while Sonbul (2015) controlled for the frequencies of the sub-parts (the adjectives and the nouns), it was not clear if all the ESL learners knew all the component words in the target adjective-noun sequences. Moreover, as Sonbul (2015) also conceded, the results from the ESL learners may have been influenced by the participants' 15 different L1 backgrounds.

The third study by Gyllstad and Wolter (2016) investigated the processing of *verb + (determiner) noun* sequences (e.g., *pay tax, break a promise*), in which the determiner was included when necessary for grammaticality. Their participants, adult native English

speakers and advanced Swedish learners of English, pressed a *YES* or a *NO* button to judge whether each word sequence on a computer screen is "meaningful and natural" (p.307) in English. Based on reaction times, phrase frequency based on lemmatized forms of the verbs[8] significantly predicted reaction times from both participant groups, regardless of whether every component word in the target phrase conveyed a literal meaning (e.g., *pay tax, rent a car*) or one of the words expressed a figurative meaning (e.g., *break* in *break a promise*).

Although these three ESL studies appeared to corroborate the ontological status of frequency effects, all or part of the stimuli consisted of two-word compositional phrases. As in L1 acquisition, stronger evidence for frequency effects would be from processing of longer compositional word sequences. Three existing studies investigated compositional three- to five-word sequences (Ellis et al., 2008; Siyanova–Chanturia, Conklin, & van Heuven, 2011; Valsecchi et al., 2013), but these studies have yielded inconclusive results and a methodological issue yet to be addressed is stimuli substring frequency control. First, a study by Ellis et al. (2008), which also included native English speakers and was therefore discussed in a previous section, investigated ESL comprehension of 108 academic word sequences. Based on reaction times in a phrasal acceptability task, the ESL learners exhibited sensitivity to phrase frequencies. Second, an eye-tracking study by Siyanova–Chanturia, Conklin, and van Heuven (2011) showed that ESL learners read binomials, or three-word phrases consisting of two content words from the same part of speech joined by a conjunction (e.g., *bride and groom*) significantly faster than less frequent reversed forms (e.g., *groom and bridge*). While the

[8] Lemmatization is the process of grouping different forms (e.g., *pay/ pays/ paying/ paid*) of the same base verb (e.g., *pay*) (Gries & Berez, 2016)

results from these two studies seem in line with usage-based accounts, the phrase frequency effects were observed in the absence of a substring frequency control.

The next study was an eye-tracking study by Valsecchi et al. (2013), who investigated frequency effects on four-word English compositional sequence comprehension in native English speakers and non-advanced ESL learners who were students at a German university. This latter group was characterized as being non-English majors who had previous experience of living in an English speaking country for less than one academic semester. The stimuli were pairs of phrases constructed from the BNC, each consisting of a higher frequency phrase (e.g., *there is no need*), and a lower frequency phrase (e.g., *there was no work*). The participants were instructed to read sentences containing the target phrases, one sentence at a time, as quickly as they can and subsequently answer a comprehension question. Based on first pass reading time and total reading time, only the native English speakers exhibited frequency effects. In this study, however, there were several factors apart from phrase frequency that could have affected the results. For example, while the researchers attempted to match the frequency of the different individual words (e.g., *need* and *work*) across all the high and low variants in each stimuli pair, the frequency of the longer subparts (e.g., *no work* vs *no need*) were not controlled for. Moreover, the target phrases were embedded in different sentences and positions (e.g., *Mr. Lumbergh says that <u>there is no need</u> to work next Saturday* vs. *Ben soon realized that <u>there was no work</u> left for his secretary*). Thus, it was not clear if or to what extent these factors affected the results.

A very recent study by Hernández et al. (2016) was the first study that demonstrated frequency effects on compositional four-word sequence comprehension in

English learners with an attempt to control for frequency of subparts of the target phrases. Using one part of the phrases in Arnon and Snider's (2010) study and a phrasal decision task, the researchers reported that phrase frequency significantly predicted reactions times in three groups of participants: 27 native English speakers, 27 adult ESL learners in the US, who were university students or college graduates from a variety of L1 backgrounds (mean length of stay in an English speaking country = 40.93 months, *SD* = 47.08), and 25 learners of English whose L1 was Spanish or Catalan and who were studying for an undergraduate translation and interpretation degree at a Spanish university (mean length of stay in an English speaking country = 2.44 months, *SD* = 2.20). Therefore, the results provided support for usage-based approaches and challenged the dual-mechanism model of language processing. It should perhaps be pointed out that the participants in the third group may be a unique group of English learners in a non-English speaking environment. That is, although they had no or little experience of living in an English speaking country, the fact that they were majoring in translation and interpretation probably means that they were very motivated and relatively proficient language learners. Indeed, based on the Lexical Test for Advanced Learners of English (Lemhöfer & Broersma, 2012) and self-reported English proficiency ratings, Hernández et al. (2016) pointed out that the English proficiency of the participants in the second and the third groups did not differ significantly. Moreover, as indicated in self-reported background questionnaires, participants in the third group had been exposed to authentic English input from various sources (e.g., classrooms, movies, songs, TV and radio programs). While the results from Hernández et al. (2016) may raise an interesting question of whether living in an English speaking country is a prerequisite for English

learners to exhibit sensitivity to compositional multi-word sequence frequency, the overall results suggested that, like native English speakers, adult English learners can also demonstrate phrase frequency effects.

Methodologically, in the previous ESL studies that reported frequency effects in adult English learners, a large native speaker corpus, such as the BNC, COCA, was the typical source of frequency data on which the target phrases were constructed. Moreover, participants' L1 may be the same (e.g., Gyllstad & Wolter, 2016; Wolter, & Gyllstad, 2013) or may differ (e.g., Ellis et al., 2008; Hernández et al., 2016; Sonbul, 2015). As a result, findings from these studies seemed to support the argument that, although the ESL input each learner receives tends to be individualized, data from a large and adequately representative corpus should represent the shared regularities of input that all language users are likely to be exposed to (Hoey, 2005; Wolter & Gyllstad, 2013).

To conclude, several previous empirical studies have investigated frequency effects in ESL compositional phrase comprehension. Most of these provided some support for usage-based approaches (Ellis et al., 2008; Gyllstad & Wolter, 2016; Siyanova–Chanturia, 2011; Sonbul, 2015; Wolter & Gyllstad, 2013), except a study by Valsecchi et al. (2013). In these studies, however, the target phrases contained two-word phrases and/ or there was no strict control of substring frequency in the analyses. These two limitations were simultaneously addressed only in a recent comprehension study by Hernández et al. (2016), which documented the effects, thus providing stronger support for usage-based approaches to L2 acquisition. Also, in the previous studies, it was not always clear if participants who were ESL learners knew all the component words in the study. If they did not know the words, especially words with low frequency, the lack of

the familiarity with component words, rather than phrase frequency alone, could have affected the results.

1.5 Previous work on the role of frequency in ESL word sequence production

There has been a scarcity of research focusing on frequency effects on ESL compositional multi-word sequence production. One related line of research consists of empirical inquiries into the pattern of verb distributions in ESL language production. As noted, one observation in L1 acquisition (e.g., Goldberg et al., 2004) is that, in child-directed and in children's speech, the distribution of verbs in a verb argument construction is Zipfian (1935). Taking this observation further and drawing on the work by Ellis and Ferreira–Junior (2009b), Ellis and Larsen–Freeman (2009) argued that the Zipian (1935) distribution is also observed in initial verb argument constructions produced by adult language learners, in the speech of their native speaker conversation partners ("foreigner talk"; Larsen–Freeman & Long, 1991), and possibly in the speech of ESL teachers. For example, Ellis and Ferreira–Junior (2009b) analyzed 5-year longitudinal data of speech produced by adult immigrants in Britain and their native-speaker conversation partners in the European Science Foundation Corpus (Feldweg, 1991). The focus was on three constructions: the verb locative (e.g., *She went to Costa Rica*), the verb object locative (e.g., *She put the cup on the desk*), and the ditransitive. The results suggested that the verbs first used by the non-native speakers in each of these constructions (*go, put, give*, respectively) were frequent in native speakers' speech, semantically prototypical, and applicable to a variety of situations. Moreover, the distribution of verbs in these constructions also seemed to be Zipfian (1935). The

observation about the Zipfian (1935) distribution in L1 acquisition has also motivated ESL psycholinguistics research using a production task, such as a task requiring participants to generate verbs that came into their mind when they saw a verb argument construction frame (e.g., *she* ____ *against the*....). The aim of such research was to analyze verb-construction associations (Römer et al., 2014).

Another related line of research consists of studies on ESL learners' production of frequent English sequences in writing (e.g., Laufer & Waldman, 2011; Nesselhauf, 2005) and speaking (e.g., Boers, Eyckmans, Kappel, Stengers, & Demecheleer, 2006; Stengers, Boers, Housen, & Eyckmans, 2011; Wood, 2006, 2010). These studies were not informed by the observation about Zipfian (1935) distribution or the prediction about frequency effects on phrase processing. Instead, motivated by the argument that an appropriate use of frequent ESL word sequences characterizes proficient L2 production (Pawley & Syder, 1983; Sinclair, 1991), the focus was on qualitative and/ or quantitative native–non-native similarities or differences in the use of frequent English word sequences or on the relationship between sequence use and ESL proficiency. Moreover, in these studies, identification of frequent word sequences, particularly in speech production, may be based on arbitrary measures, such as judgment of a few native or non-native English teachers with many years of English teaching experience (Boers et al., 2006), a few native English speakers with a linguistics background and familiarity with phraseology (Stengers et al., 2011), or the researcher's own criteria (Wood, 2006, 2010). These measures were therefore not corpus-derived statistics measures. A study by Schmitt, Grandage, and Adolphs (2004) differed from the other ESL production studies in that it looked at ESL learners' oral production of corpus-based frequent word

sequences (e.g., *as a consequence of, for example*) to investigate whether these sequences were stored holistically. Because the learners did not necessarily produce these sequences fluently and without errors, the researchers suggested that these sequences were not holistically stored.

However, these two lines of ESL production research did not seek to investigate phrase frequency effects based on measures which presumably tap into the strength of relations between words in compositional phrases, such as phonetic reduction or phrase production durations, which were investigated in the L1 literature (e.g., Bannard & Matthew, 2008; Bybee & Scheibman, 1999). To date, I am aware of only one ESL study which focused on frequency effects on compositional multi-word sequence production, that is, the study by Ellis et al. (2008). As discussed in a preceding section, Ellis et al.'s (2008) stimuli consisted of 108 academic word sequences. To investigate frequency effects, the researchers conducted two elicited oral production experiments. In the first experiment, the target phrases were presented visually, one phrase at a time, and participants read the stimuli as fast as they could after each phrase appeared on a computer screen. VOT, the duration between the onset of the stimuli and the onset of their response, and speech duration, or the duration between the onset and offset of articulation, were the outcome variables. In the second experiment, participants were presented with a preceding part of a word sequence (e.g., *a wide variety*) and pressed the space bar to make it disappear, and then saw the final word of the sequence (e.g., *of*), which they had to read aloud. The outcome variable was voice onset time of the final word (final word VOT), or the time between the onset of the final word and the beginning of the participant's speech response. The ESL learners in the two experiments

were six and 16 international students respectively. They were from a variety of L1 backgrounds and were enrolled in an English for academic purposes course at a US university. Results from native English speakers revealed that phrase frequency significantly predicted VOT and final word VOT, but not speech durations. In case of the ESL learners, frequency significantly predicted VOT and almost significantly predicted speech durations. While overall these results lent some support to the psychological reality of frequency effects, as noted, in this study the frequency of subparts of the target sequences were not strictly controlled for. Consequently, this lack of substring frequency control may be responsible for the absence of frequency effects on the native speakers' speech durations—a possibility empirically supported by Arnon and Priva (2013). Certainly, the lack of such control may have influenced the results from the ESL learners as well. Moreover, the stimuli in Ellis et al.'s (2008) study were constructed based purely on corpus-based statistics measures. Some target phrases were complete syntactic constituents, such as *a great deal of* (a determiner phrase), but others were not and did not seem to be intonational phrases (e.g., *and at the, and so on but, the way in which the*). The participants also had a variety of L1 backgrounds. These cast doubt on whether phrase frequency was the sole determinant of production durations.

In sum, previous studies with native English speakers have demonstrated frequency effects on compositional multi-word sequence recognition (Arnon & Snider, 2010; Hernández et al., 2016) and production (Arnon & Priva, 2013, 2014; Bannard & Matthew, 2008) when frequency of the subparts of the target sequences was controlled for. The results seemed to support usage-based researchers' prediction and have implications that language comprehension and production models should accommodate

frequency effects beyond the bigram level. Despite the possible differences in L1 and L2 acquisition, previous ESL comprehension studies have lent support at varying degrees to frequency effects on different types of compositional ESL sequences, but so far only the recent study by Hernández et al. (2016) has provided relatively stronger support for usage-based researchers' claim because it investigated compositional sequences beyond the bigram level and controlled for substring frequency of the test phrases. Also, in the previous ESL studies, it was not always clear if the learners knew all the component words in the target phrases, and thus participants' lack of familiarity with component words, instead of phrase frequency alone, could have influenced the results. Moreover, as Siyanova–Chanturia and Martinez (2014) also argued, to date, little is known about L2 compositional online multi-word sequence production. While the study by Ellis et al. (2008) has suggested that frequency effects may be observed in L2 production, more research in this area is much needed. Further, there has been no research which simultaneously looks at frequency effects on language comprehension and production in the same participants. Such research may reveal if task characteristics (comprehension vs production) could impact frequency effects.

1.6 Research questions in the current study

The research questions that guided the current study were as follows.

1. Are adult native English speakers and ESL learners sensitive to the frequency of compositional four-word phrases in recognition when the frequency of the smaller parts is controlled for?
2. Are adult native English speakers and ESL learners sensitive to the frequency of

compositional four-word phrases during language production when the frequency of the smaller parts is controlled for?

Two respective experiments, discussed in the next two chapters, were conducted to answer these questions.

CHAPTER 2: EXPERIMENT I

2.1 Research question and prediction

This experiment was conducted to answer the first research question; the aim was to test frequency effects on the processing of compositional multi-word sequences processing in language comprehension when the frequencies of the sub-parts of these sequences are controlled for. In light of previous research (e.g., Arnon & Snider, 2010; Wolter & Gyllstad, 2013), receptive processing was operationalized as reaction time in a phrasal acceptability task, in which participants decided whether target phrases are possible English word sequences, and sequences that are more frequent, such as *don't have to worry*, should be encountered more often and therefore should be recognized and comprehended faster than less frequent phrases, such as *don't have to wait*. Participants in this experiment were adult native English speakers and ESL learners. Given the results from the two previous compositional multi-word phrase comprehension with substring frequency control (Arnon & Snider, 2010; Hernández at al., 2016), I expected that frequency effects should be observed from the native English speakers. As for the ESL learners, given Hernández et al.'s (2016) recent findings, these learners could also exhibit the effects. Results from the current experiment could therefore support usage-based approaches (e.g., Bybee, 2010; Ellis, 2002), and counter the prediction in the dual mechanism model, which posits that only individual words, bound morphemes, and idioms are susceptible to frequency effects (Prasada et al., 1990; Prasada & Pinker, 1993; Ullman, 1999). The results could also counter the claim that ESL learners cannot retain memory about L2 word co-occurrences (Wray, 2002).

2.2 Method

2.2.1 Participants

Participants were a sample of 51 adult native English speakers (Male = 19, Female = 32) and 52 adult ESL learners (Male = 18, Female = 34). The first group consisted of undergraduate and graduate students enrolled at a large Midwestern US university. Their average age was 20.58 years old (Min = 18, Max = 33, *SD* = 2.93). In both groups, participants had a variety of educational backgrounds, such as engineering, business, international relations, chemistry, nursing, and advertising. None majored in linguistics or had a hearing impairment or speaking difficulty. These participants also participated in the second experiment and received compensation only upon the completion of the second experiment.

The ESL learners in the second group were international undergraduate or graduate students enrolled at the same university. Their characteristics based on the background questionnaires were summarized in Table 3.[9] The average age at which they started learning English was 9.65 years old. If any of them reported beginning learning English at an age lower than 10, I additionally asked if they had any immersion experience (e.g., in an English immersion school) in their early years. None of them reported having such experience—they began learning English in a formal English classroom in their home country. As a result, the fact that some of these participants began learning English at a young age (i.e., below 10) should not make these participants differ significantly from the other participants in this group. As Muñoz (2008) observed, to date no evidence has suggested that an earlier start of English learning in a non-

[9] More details about this questionnaire are in section 2.2.2.3. The questionnaire is shown in Appendix E.

immersion environment, in which the English input is less intense than input in an immersion environment, guarantees higher ESL proficiency given the same amount of instructional time. In fact, in a non-immersion environment, even younger starters who have received more instruction do not necessarily demonstrate the benefit of beginning learning English at a younger age. Moreover, all the ESL learners shared the same L1, which was Chinese. Because the task in the first and second experiments involved English reading, this L1 control eliminated the possible effect of L1–L2 orthographic differences in reading that can arise due to participants' dissimilar L1 backgrounds (Grabe, 1991).

Table 3. ESL learners' background (*N*=52)

	Maximum	Minimum	*M*	*SD*
Age	32	18	23.54	3.93
TOEFL score	109	85	95.52	6.63
Age at beginning English instruction	16	5	9.65	2.63
Years of residence in the US	3.5	2	2.61	0.56
Self-rated ESL skills (out of 5)				
Speaking	5	2	3.73	0.73
Listening	5	2	4.14	0.71
Reading	5	2	4.00	0.82
Writing	5	2	3.72	0.98
Self-reported percentages of daily English use				
Speaking	98	20	54.90	22.06
Listening	98	30	57.47	19.19
Reading	98	25	67.10	19.47
Writing	100	30	76.31	18.02

Note. Percentages of daily use of English was from 51 ESL learners because one learner did not show up for the second experiment and thus did not complete the part of the background questionnaire which asked about this topic.

Moreover, to control for ESL proficiency and to increase the chance that the ESL learners were familiar with the words in the target phrases, I recruited only learners who

had a minimum internet-based TOEFL score of 85 out of 120 or a minimum IELTS score of 6.5 out of 9. International university students with this score or higher pass the English proficiency requirement for undergraduate admissions to many US universities. As a result, the ESL learners in the current study can be characterized as being proficient enough to study in an English-speaking environment. The minimum score was slightly higher than the minimum score required in several previous studies on ESL word sequence memory retention or comprehension (e.g., Durrant & Schmitt, 2010; Jiang & Nekrasova, 2007; Sonbul, 2015), in which participants had a minimum paper-based TOEFL score of 550, roughly equal to 80 in the internet-based test, or a minimum IELTS score of 6. Among the 52 ESL learners, only two took the IELTS[10] instead of the TOEFL and received scores of 7.0 and 7.5. According to the Educational Testing Services, these IELTS scores were roughly equivalent to the internet-based TOEFL scores of 95 and 100 respectively[11]. The group's average TOEFL score, including the two converted scores, was 95.52.

Because ESL processing can also be influenced by the amount, the quality, and the structure of previously–encountered ESL input, as well as the recency of ESL exposure (Ellis, 2002; Littlemore, 2009; Muñoz, 2008), I attempted to minimize the influence of these factors. First, the recruited ESL learners had lived in the US for approximately 2-3 years and reported no previous experience of living in another English speaking country prior to coming the US, except for one participant who had lived in Canada for five months. Therefore, the ESL learners should have a relatively similar

[10] The academic program these participants were enrolled in accepted either TOEFL or the IELTS scores.
[11] https://www.ets.org/toefl/institutions/scores/compare/

length of stay in an English-speaking environment. Second, since all these learners were in the US at the time of this study, recency of English exposure should not have been a major factor contributing to differences in their ESL phrase processing. However, the attempt to control for these variables was by no means perfect. For example, as Table 3 shows, the self-reported average percentage of time these participants listened to English on a daily basis at the time of the study ranged from 30% to 98%. It was also possible that these learners travelled out of the US during the 2-3 years period, and they probably had had a different amount of exposure to English input (e.g., through TV programs or movies) before coming to the US. The length-of-stay criteria may therefore not perfectly reflect the amount of English input they had been exposed to up to the time of this experiment. However, the attempt to control for these variables should be reasonable given the practical considerations of participant recruitment.

The native English speakers were recruited through flyers posted on the university's campus. Some ESL learners were also recruited through on-campus flyers, but many were recruited through online student communities, a university student mailing list, and my personal connections. I also submitted a request to the university's office of the registrar to help distribute a recruitment email to Chinese-speaking students who passed the minimum TOEFL requirement of this study. This helped recruit approximately 30% of the participants in this group.

2.2.2 Materials

2.2.2.1 Target phrases

I used a part of the materials from Arnon and Snider (2010). The target phrases were 28 pairs of phrases. Each pair consists of two four-word phrases that differed in

phrase-frequency (high vs low) and differed only in the final word (e.g., *don't have to worr*y vs *don't have to wait*). Consequently, the first three words were matched for frequency. The two variants in each pair were also of the same constituent types (e.g., verb phrases, noun phrases). These target pairs were constructed from the Switchboard (Godfrey et al., 1992) and Fisher (Cieri et al., 2004) corpora, which were combined to form a 20-million-word corpus. According to Arnon and Snider (2010), these corpora of American English telephone conversations were chosen because the corpora could be used to create target phrases which are common in natural English conversations and generally could form an intonational phrase. Moreover, these corpora were not based on a specific genre (e.g., articles in the Wall Street Journal) which may not be common to all English speakers. I downloaded the corpus text files from the Linguistic Data Consortium (https://www.ldc.upenn.edu) and derived frequencies of the phrases and their subparts with TextWrangler (http://www.barebones.com/products/TextWrangler/), a free software program which is able to handle a large amount of text data.[12]

In the current study, it was important that the participants who were ESL learners knew all the component words in the target phrases. Otherwise, it would be unclear whether their ESL phrase processing was affected by phrase frequency alone, or also by lack of familiarity with the component words. Based on their English proficiency and their experience of living in the US, the recruited ESL learners should be familiar with the component words. However, to further ensure this familiarity, before the present experiment was conducted, I entered the words into the Oxford 3000 Word Checker[13],

[12] I am thankful to Dr. Suzanne Wagner for her suggestions about the Linguistics Data Consortium and TextWrangler.
[13] http://www.oxfordlearnersdictionaries.com/us/oxford_3000_profiler.html

which reported that all the words are on the Oxford 3000 list of most important and frequent English words. According to Oxford, if 100% of words in an English text are on the list, the text should be within the capacity of low intermediate ESL learners. As a result, the participants should be familiar with the component words. In addition, at the end of the second experiment, I presented a list of the target phrases to the ESL learners and asked them to identify any component words they did not know. They reported that they knew all the words. Thus, in the present study, participants' phrase processing should not have been influenced by a lack of familiarity with component words.

The four-word phrases in the 28 pairs were selected by Arnon and Snider (2010) based on four criteria. First, the first three words (e.g., *don't have to*) had a high frequency of at least 30 per million words in the 20-million word corpus. Second, the last word, or the word that differed in the high and low frequency phrases in each target pair, had to occur at least 50 times per million words. Arnon and Snider (2010) used these two criteria to increase the reliability of the frequency estimates for the low frequency phrases. Using TextWrangler, I obtained frequency data from the corpus myself and found that these two criteria were indeed met. Third, the last word in the target phrases was not a determiner, which would create an incomplete intonational phrase. Therefore, the target phrases did not include a four-word phrase such as *I talked to the (professor)* or *I have bought that (book)*. Finally, the last word in the target phrases was not a demonstrative pronoun (e.g., *that*), which can also be interpreted as a determiner.

The 28 target pairs were from a high- and a low-cutoff bin, which consisted of 16 and 12 target pairs respectively.[14] Table 4 shows examples of the phrases. In each pair,

[14] These were the stimuli in the first experiment in Arnon and Snider's (2010) study.

one phrase was classified as a high frequency phrase and the other was classified as a low frequency phrase. The first two pairs are examples from the high cut-off bin. In the original study, the cut-off point for classifying a phrase as a high or a low frequency phrase in this bin was 10 occurrences per million. However, based on the frequencies I derived from Fisher and Switchboard, each high frequency variant occurred at least 12.0 times per million words, while each low frequency variant occurred less frequently.[15] The third and fourth pairs are examples from the low cut-off bin. As in the original study, the cut-off point for classifying a phrase as a high or a low frequency phrase was one occurrence per million. That is, in each pair, the high frequency variant occurred at least one time per million words, while the low frequency variant occurred less often. Thus, the classification of a phrase as having high or low frequency was meaningful within each cut-off bin, not across all the 28 stimuli pairs.

Table 4. Examples of the 28 target pairs

Cut-off bin	Phrases	Frequency condition	Frequency (per million words)
High	Don't have to worry	high	20.4
	Don't have to wait	low	2.0
	I don't know why	high	47.9
	I don't know who	low	11.6
Low	A lot of rain	high	6.0
	A lot of blood	low	0.3
	Don't have any money	high	2.8
	Don't have any place	low	0.5

The purpose of incorporating the phrases from the two cut-off bins in the current study was twofold. First, Arnon and Snider (2010) observed frequency effects from adult

[15] I use a decimal point because, based on the frequencies I could derive, one low frequency variant has a frequency of around 11.6 occurrences per million words, while one high frequency variant has a frequency of 12.0 occurrences per million words.

native English speakers in a phrasal acceptability task in both cut-off bins. In light of the two proposals about phrase frequency representation I discussed in Chapter I and their results, Arnon and Snider (2010) argued that there was no direct evidence that highly frequent phrases (i.e., those with frequency above the cut-off point in the high cut-off bin) were stored as a whole and were processed faster than phrases with lower frequencies. Instead, because frequency effects were observed in both cut-off bins in a similar way, Arnon and Snider (2010) argued that the differences in reaction times to the target phrases should have resulted from relative quantitative differences (i.e., different frequencies of previous phrase encounters) (Bybee, 2006, 2010; Bybee & Hopper, 2001). In this experiment, I therefore investigated whether Arnon and Snider's (2010) results are replicable. The second reason for having the two cut-off bins was specifically related to the E"SL learners, who presumably had had less exposure to English input. These learners may exhibit frequency effects only with target phrases in the high-cut off bin because they may have been exposed to many instances of high frequency phrases. On the other hand, all except one high frequency phrase in the low-cut off bin had frequencies below the cut-off point in the high cut-off bin (i.e., below 12.0 occurrences per million words). It may be possible that the ESL learners had not had much exposure to the target phrases in the low cut-off bin and thus may not demonstrate frequency effects in this bin. In short, an additional goal of having the two cut-off bins was to illuminate whether ESL learners had stored sufficient accumulated statistics information (i.e., frequency of occurrences) through previous ESL exposure to demonstrate frequency effects in both cut-off bins. The 28 target pairs and their frequencies are listed in Appendix A.

In the high cut-off bin and based on the frequency data I derived from Fisher and Switchboard, the mean frequencies of the high and low frequency variants across the 16 target pairs were 25 occurrences (Min = 12.00, Max = 53.15, *SD* = 12.97) and 4.87 occurrences (Min = 0.70, Max = 11.60, *SD* = 3.93) per million words respectively. Across these pairs, frequencies between high and low frequency phrases differed significantly, t (30) = -10.76, $p < .001$. In the low cut-off bin, the mean frequencies of the high and low frequency variants across the 12 target pairs were 4.68 occurrences (Min = 1.85, Max = 12.60, *SD* = 3.18) and 0.27 occurrences (Min = 0.05, Max = 0.55, *SD* = 0.14) per million words respectively. Across these 12 pairs, frequencies between high and low frequency phrases also differed significantly, t (22) = -4.81, $p < .001$. In sum, across the target pairs in each cut-off bin, high frequency phrases occurred significantly more often than low frequency phrases.

Regarding the sub-parts of the target phrases, because the first three words in each pair were identical (e.g., *don't have to worry* vs *don't have to wait*), each pair differed only in three subparts: the last word (*worry* vs *wait*), the last two words (e.g., *to worry* vs *to wait*), and the last three words (e.g., *have to worry* vs *have to wait*). In the high cut-off bin, between the high and low frequency variants across the 16 pairs, there was no significant difference in the frequencies of the last words, t (30) = -0.24, $p = .813$, the last two words, t (30) = -1.03, $p = .310$, and the last three words, t (30) = -0.95, $p = .350$. Similarly, in the low cut-off bin, between the high and low frequency variants across the 12 pairs, there was no significant difference in frequency of the last words, t (22) = -1.15, $p = .264$, and the last two words, t (22) = -1.68, $p = .117$. There is however a significant difference in the frequency of the last three words, t (22) = -2.95, $p = .013$, with the last

three words in higher frequency phrases being more frequent. Therefore, to strictly control for the frequencies of the three sub-parts across the target pairs and to observe effects of the four-word sequence more clearly, the frequencies of these subparts were later entered as control variables in the analysis. The purpose was to ensure that any difference between the participants' processing of high and low frequency variants in each bin can be attributable to the difference in whole phrase frequency, not to the difference in subpart frequency. The frequencies of the subparts of the target phrases are in Appendix B.

With regard to the corpora, by using Fisher and Switchboard, I attempted to replicate results on part of the native speakers in Arnon and Snider's (2010) study and take advantage of the stimuli substring frequency control. However, because several previous ESL studies on word sequence comprehension used frequency data obtained from a larger corpus, such as COCA (e.g., Kim & Kim, 2012; Wolter & Gyllstad, 2013), I investigated the correlation between frequency of the target phrases obtained from the combined Fisher and Switchboard corpus and frequency derived from COCA. Based on the Kolmogorov–Smirnov tests, log frequencies of the 28 target pairs were normally distributed (*ps* = .200), regardless of whether the frequencies were derived from (1) the combined Fisher and Switchboard corpus, (2) the spoken portion of COCA, which contained more than 79 million words from conversations in American TV and radio programs, or (3) the whole COCA corpus, which, besides the spoken data, contained written texts from various sources (e.g., magazines, academic journals, fiction; Davies, 2013).[16] Log frequencies of the target phrases derived from the first source correlated

[16] Log transformation reduces skewness of data (Field, 2009).

significantly with those from the second source, $r = .49$, $p < .001$, and third source, $r = .43$, $p < .001$. However, based on frequencies obtained from the second and third sources, in four target pairs, high frequency variants (i.e., *a lot of rain*, *you like to read*, *don't know how much*, and *we have to talk*) had lower frequencies than their low frequency counterparts (i.e., *a lot of blood*, *you like to try*, *don't know how many*, and *we have to say*, respectively). Therefore, generally frequency data of the target phrases from Switchboard and Fisher combined correlated well with frequency data from COCA, despite some differences among these sources. Also, unlike some previous related ESL research, this study did not include frequency data from a written component of a native speaker corpus. The use of spoken corpora in the current study should be compatible with the discussion about L2 acquisition in usage-based approaches, which has been informed by the child L1 acquisition literature (e.g., Goldberg et al., 2004; Tomasello, 2003, 2009). As discussed in the literature review, so far there has been no direct discussion of the role of written input in usage-based L2 acquisition, despite a possibility that such input may play a role.

Finally, in terms of the meaning of the target phrases, Arnon and Snider (2010) asked 25 native English speakers to rate the plausibility of the low and high frequency phrases. The researchers reported that, based on Wilcoxon tests, there were no significant differences in the rating scores between high frequency phrases and low frequency phrase in the high cut-off bin (M for high frequency variants = 6.7; M for low frequency phrases = 6.7; W = 113.5, $p > .5$) and the low cut-off bin (M for high frequency variants = 6.6; M for low frequency phrases = 6.6; W = 43.50, $p > .5$). These were not surprising because all the phrases are possible and meaningful English phrases.

As a result, any differences in processing between high and low frequency phrases in each cut-off bin should not have resulted from a difference in the plausibility of the meaning conveyed by the phrases.

2.2.2.2 Fillers

In addition to the 56 target phrases, there were 80 four-word-phrase fillers of two types. As in Arnon and Snider's (2010) study, the first were 12 possible English phrases (e.g., *hold a green bag*), and the second were 68 impossible English phrases. Among the latter, 75%, or 51 phrases, had a wrong word order (e.g., *girl the was sad*), while 25%, or 17 phrases, were impossible due to an incorrect preposition use (e.g., *look with the sky*). An attempt was made to avoid an overlap between words in the target phrases and the fillers. Therefore, in this experiment, there was an equal number of possible phrases (56 target phrases plus 12 fillers) and impossible phrases (68 fillers). A part of these fillers were from Arnon and Snider's (2010) study[17], but I also constructed additional fillers to obtain the target number. The Oxford 3,000 Word Checker indicated that 98% of the words in the fillers were among the most frequent English words, so participants who were ESL learners should understand the words in the fillers, listed in Appendix C.

2.2.2.3 Background questionnaires

For participants who were native English speakers, I created a questionnaire which asked about their personal and educational backgrounds. Another questionnaire was constructed for the ESL learners, and additionally asked about the age when they

[17] I thank Inbal Arnon for sending me examples of the fillers she used in the original study.

started learning ESL, their daily use of or exposure to English, and their previous experience of living in an English-speaking country. Because frequency effects result from accumulation of statistics in previously-encountered input, one question asked whether they had a record of hearing impairment, which may have influenced the amount of input they had received. Moreover, since the second experiment involved speech production, there was a question as to whether the participants had a record of speech production difficulty. The questionnaire for each group was divided into two parts, to be filled during the break in the first and the second experiments. The questionnaires for the native English speakers and ESL learners are shown in Appendices D and E respectively. Some of the questions in these two questionnaires were adapted from those used by Marian, Blumenfeld and Kaushanskaya (2007).

2.2.3 Procedure

I ran the experiment on Superlab (Cedrus Corporation, 2006) and conducted the experiment with one participant at a time. Participants sat in a quiet room in front of a computer screen and completed a phrasal decision task, in which they saw four-word phrases in the center of the screen and judged whether the phrases were possible English word sequences. Each participant completed two experiment blocks. The procedure in this experiment is illustrated in Figure 1.

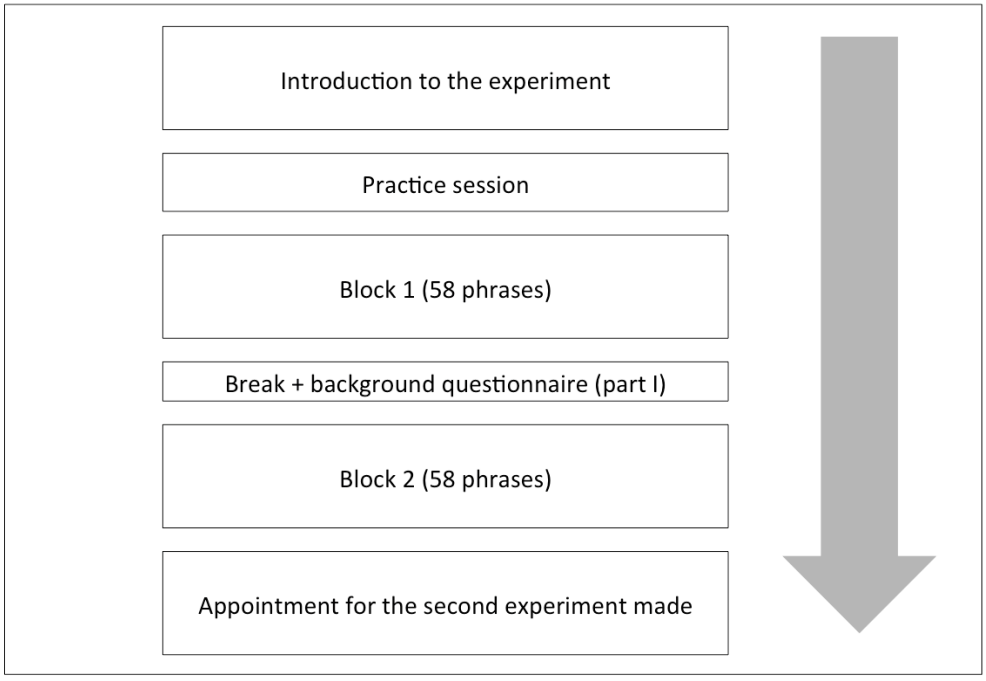

Figure 1. Procedure in the phrasal acceptability judgment task

Before the experiment began, each participant read and signed a consent form. At this point, the participant was merely informed that the objective of the current study was to investigate English phrase comprehension and production. The participant then sat in front of a computer on which the experiment was run, and was presented with the following instruction in Figure 2.

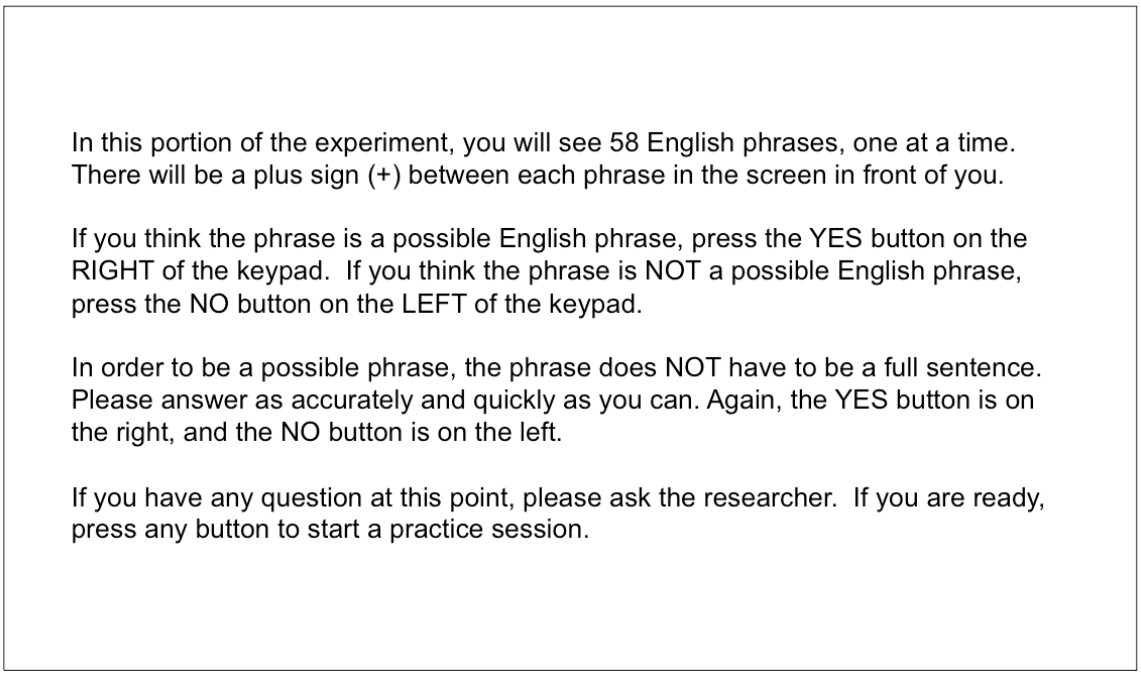

Figure 2. Screen shot of instruction in the phrasal acceptability judgment task

These instructions were for most participants, who were right-handed. The *YES* and *NO* buttons were equally positioned on the keypad. I provided different instructions to left-handed participants: to press the *YES* button on the left and *NO* button on the right. Thus, for left-handed participants, a different version of the keypad was used, but the *YES* and *NO* buttons were also equally positioned on the keypad.

At the beginning of both experiment blocks, participants completed a short practice section, in which they saw examples of both possible and impossible sequences. This section, included to familiarize the participants with the task, contained six practice items. The impossible sequences contained the same type of errors (i.e., wrong word order or incorrect preposition use) that the ungrammatical phrases in the actual experiment contained. Participants were also allowed to ask any question that they may have after the practice section, but were informed that they cannot ask any question when

the actual experiment began. I also instructed all the participants to continue the experiment even when they felt they made a mistake.

The presentation of phrases is illustrated in Figure 3. As in previous research (Tremblay & Tucker, 2011; Wolter & Gyllstad, 2013), participants first saw a plus sign in the center of the screen for eye fixation. In this experiment, the sign lasted just 333 milliseconds and was followed by a blank screen, which lasted for 50 milliseconds. The phrase then appeared and remained on the screen until a button press. The phrases appeared one at a time and in their entirety (font: Arial; size: 36; position: center). Words in the phrases, except the first person personal pronoun and proper names, were in the lower case. The outcome measure was their reaction time, or the duration between the onset of the phrase and a button press, but their judgment accuracy was also recorded.

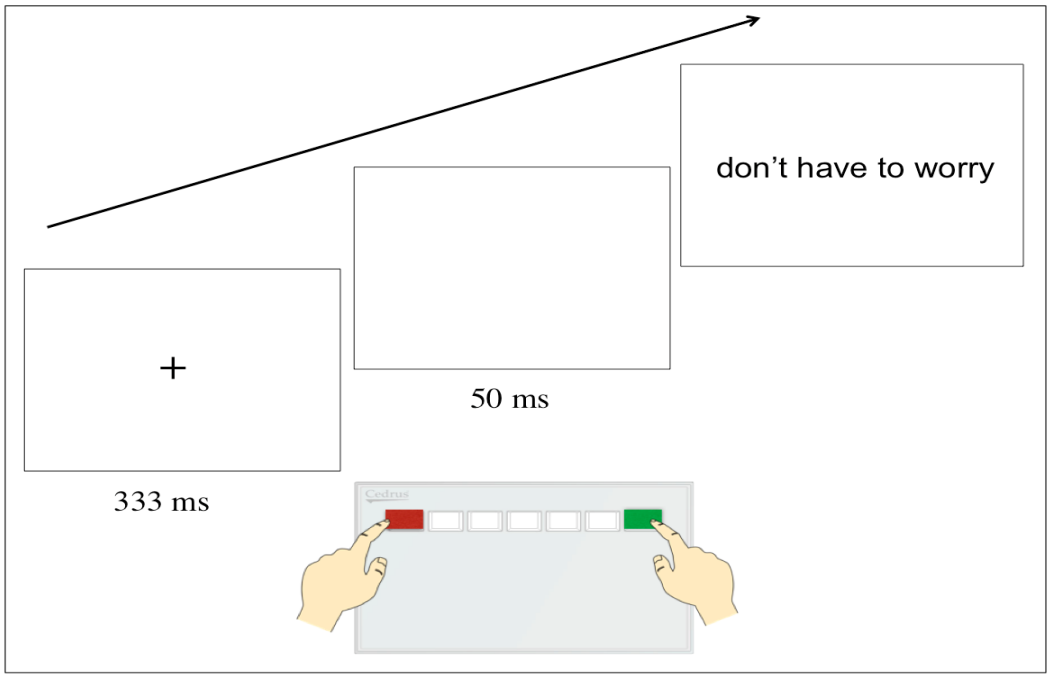

Figure 3. Phrase presentation in the phrasal acceptability judgment task

The target phrases and the fillers were divided into two blocks: A and B. One variant from each of the 28 target pairs was randomly assigned to only one block. The purpose was to minimize a possible repetition effect resulting from participants seeing the identical first three words in the two variants of the same target pair (e.g., *don't have to worry* and *don't have to wait*) in the same block. Each block thus consisted of 14 high frequency variants and 14 low frequency variants from the target pairs. Fillers were randomly and equally assigned to each block such that, in total, each block contained 28 target phrases (14 high frequency variants and 14 low frequency variants from the target pairs), six fillers which are possible English word sequences, and 34 fillers which are impossible sequences. The total number of phrases in each block was thus 68, with half of the phrases being grammatical and the other half ungrammatical. The stimuli were presented in a random order.

This experiment followed a within-subject counterbalanced design. Half of the participants in each group (native English speakers and ESL learners) were randomly assigned to complete block A first, while the other half completed block B first. Each participant completed both blocks; therefore, they saw the two variants from each target pair across the two blocks. Between the two blocks, there was a break in which participants completed the first part of their background questionnaire. The break was included to further reduce the possible repetition effect. The whole first experiment took approximately 20 minutes. At the end, I scheduled a date and time for the second experiment with each participant.

2.3 Analysis

The output files from Superlab were in a format that can be readily converted into a format for the analysis. Because the frequency cut-off point in each stimuli cut-off bin was different, I ran a separate analysis for each cut-off bin, including only reaction times for the target phrases (i.e., reaction times for the fillers were excluded). Among the 51 participants who were native English speakers, I removed one participant due to a relatively low level of judgment accuracy of 88%. The remaining 50 native English speakers (Male = 18, Female = 32) had a mean accuracy rate of 98% (Min = 91%, Max = 100%, *SD* = 0.02). Moreover, I excluded two ESL learners because they occasionally stopped during the experiment to do activities not related to the experiment (e.g., looked at their mobile phone screen); this could have affected the reaction times. Another ESL learner was excluded because of a low accuracy rate of 77%. The remaining 49 ESL learners (Male = 17, Female = 32) had a mean accuracy rate of 97% (Min = 89%, Max = 100%, *SD* = 0.03). Therefore, participants in both groups did not seem to have any difficulty doing the task. I additionally eliminated incorrect responses from the analysis. In each group, no participant was removed as an outlier because, in each frequency condition in each cut-off bin, no participant had a mean reaction time that fell outside +/- 2 *SD*s from the group mean. However, in each group, reaction times exceeding +/- 2 *SD*s from the group mean in each frequency condition in each cut-off bin were excluded. This resulted in a removal of 3% and 4% of the correct responses from the native speakers and the ESL learners respectively. There were no reaction times below 200 milliseconds (e.g., Hernández et al., 2006), which may indicate an accidental button press.

Table 5 lists all the predictors of the participants' reaction times in the current experiment. Of main interest were frequency condition of the target phrases and the participant group. In light of the findings from previous research (Hernández et al., 2016), the ESL learners should have slower reaction times than native English speakers. In the analysis, the interaction between frequency condition and group was also investigated. The other explanatory variables were control variables, entered to account for other differences between the two phrases in each pair that may have affected participants' reaction times. These included the block order (i.e., whether the participant completed that block as the first or the second block) and the number of characters of the target phrases. Based on previous research (Arnon & Snider, 2010; Hernández et al., 2016), the participants in both groups were likely to have shorter reaction times when they did a second experimental block, whether it was block A or B, due to greater task familiarity. Also, participants' reading and thus reaction times were likely to be longer when the number of characters in the target phrases increased. The other three control variables were the frequencies of the subparts that differed in each pair. As discussed, entering these frequencies was necessary—this experiment investigated whether the frequency condition of the whole target phrases significantly predicted reaction times when the frequencies of these subparts were controlled for. Following previous research (e.g., Arnon & Snider, 2010; Sonbul, 2015), I also did a log transformation of the dependent variable and the frequencies of these subparts to reduce skewness of the data.[18] Moreover, for categorical variables, one level, marked with an asterisk in the Table 5, was assigned the reference or baseline category. To illustrate the magnitude of the

[18] In the current study, the transformation was based on the base-10 logarithm.

explanatory variables, I also standardized continuous predictors. The coefficients of the standardized predictors thus indicated the change in reaction times on the logarithmic scale associated with a one standard deviation change in these predictors.

Table 5. Explanatory variables in the first experiment

Variable	Type	Level
Phrase frequency condition	Categorical	low*, high
Participant group	Categorical	NS*, ESL
Block order	Categorical	first*, second
Number of characters of the whole phrase	Continuous	
Log frequency of last word	Continuous	
Log frequency of last two words	Continuous	
Log frequency of last three words	Continuous	

Note. NS = native English speakers; ESL = ESL learners. An asterisk marks the reference category for each categorical variable.

For the analysis, I used mixed-effects regression modeling (Baayen, Davidson, & Bates, 2008; Bates, 2010). Models were run in R, a language and environment for statistical analyses (R Core Team, 2015), with the statistics package *lme4* (Bates, 2010; Bates, Mächler, Bolker, & Walker, 2015). This type of models consists of two types of effects: fixed and random. As Winter (2013) pointed out, fixed effects are associated with parameters of explanatory variables typically included in a simple linear regression model. These variables can be either continuous variables or categorical variables that a researcher manipulates or controls. Such variables represent the systematic part of a regression model (i.e., as opposed to the stochastic part or the error term), and an estimation of the parameters of these variables is of interest to the researcher. In the current experiment, therefore, variables associated with fixed effects were phrase frequency condition, participant group, block order, the number of characters in the target phrases, and frequencies of the three subparts.

On the other hand, random effects are associated with variables whose effects constitute variability specific to the randomly selected sample. In linguistics experiments, these variables typically include *subject* and *test item* (Baayen, 2008; Baayen et al., 2008; Winter, 2013). Baayen et al. (2008) explained that *subject* is associated with a random effect because a dependent variable may be affected by various factors related to individual differences (e.g., genetic, developmental, social, or even chance factors) specific to the randomly sampled participants. For example, a participant in the sample may have a higher or a lower mean reaction time compared to other participants with the same characteristics (e.g., L1, age, or gender) due to a random factor. According to Baayen et al. (2008), a similar logic can be applied to *test item*. That is, stimuli in a linguistics experiment do not encompass all possible syllables, words, phrases, or sentences in a language, and there may be random factors about each test item that affects the dependent variable. This is also the case in the current study—the 56 target phrases were randomly sampled, four-word compositional phrases out of all the possible four-word compositional English phrases, and there may be some random idiosyncrasy specific to each of these phrases that affected the participants' reaction times. Winter (2013) further notes that taking into account by-subject and by-item variability essentially gives some additional structure to the stochastic part or the error term in a regression model, although the resulting model still has remaining errors due to other factors that the research does not or cannot control. Moreover, as Baayen et al. (2008) pointed out, accommodating both fixed and random effects constitutes the most important advantage of mixed effects models because such accommodation "allow[s] the researcher to simultaneously consider all factors that potentially contribute to the

understanding of the structure of the data" (p.409). Further, Baayen et al. (2008) pointed out that because an aim of a quantitative study is to make generalizations to a larger population, results from a mixed effects model allow researchers to distinguish between the variance in the dependent variable that is explained by fixed effects, which can potentially be generalized to the larger target population, and random effects specific to randomly sampled participants and test items.

A major reason for the use of a linear mixed-effect regression model in the present study lies in its efficiency and flexibility. In ANOVA, multiple observations from each subject need to be averaged for a subject analysis, and multiple observations for each test item need to be averaged for an item analysis. The purpose of doing this is to avoid a violation of the assumption of independence. That is, multiple responses from the same participant are not independent from each other, and neither are multiple responses to the same test item. Without such averaging, the chance of a Type I error in statistical analyses seriously increases (Field, 2009). However, item-variability is disregarded in a subject analysis, and subject variability is disregarded in an item analysis, and the separation of subject and item analyses (F_1 and F_2 analyses) has both advantages and disadvantages (e.g., Baayen et al., 2008; Barr, Levy, Scheepers, & Tilly, 2013). By contrast, a mixed-effects regression model allows for cross-random effects of subjects and items. It takes into account every single observation and simultaneously accommodates both subject and item variability, allowing each subject and each test item to have a different mean reaction time in the same model (Baayen, 2008; Baayen et al., 2008; Winter, 2013). That is, it accommodates the two types of random effects discussed in the preceding paragraph. Due to its advantages, mixed-effects modeling has been

common in various academic disciplines—such as science, medicine, and linguistics (e.g., Baayen et al., 2008; Faraway, 2006; Hout, Fox, & Muniz–Terrera, 2015)—and has recently gained increasing attention in L2 acquisition research (e.g., Cunnings & Finlayson, 2015; Godfroid & Uggen, 2013; Gyllstad & Wolter, 2016; Sonbul, 2015; Wolter & Gyllstad, 2013). In the current experiment, multiple reaction times were observed from each participant and multiple reaction times were observed from each target phrase. I therefore used this mixed-effects regression modeling to take advantage of its power and accuracy.

Following the literature (Fox, 2008; Fox & Weisberg, 2011), I used Wald statistics to obtain p-values and determine the significance of explanatory variables from mixed effects regression models. Given the fixed effects in my model—which were either categorical variables with only two levels or were continuous variables—the Wald Statistics for each predictor has a chi-square distribution with a degree of freedom equal to one. Type II sums of squares was used; therefore, the significance of each predictor was assessed when all the other predictors were simultaneously controlled for, and the order of the predictors in the regression model did not matter. This allowed for a meaningful interpretation of the results because the aim of the current experiment was to see the effect of a predictor (e.g., frequency condition) when the other predictors (e.g., block order, participant group) were simultaneously taken into account. Moreover, another source of flexibility of a mixed effect regression model is that it allows a researcher to investigate whether an explanatory variable of interest has a different effect on each individual participant. In the context of the current study, this means that it allows frequency to have different effects on reaction times from each participant (i.e.,

different reaction time slopes for the effect of frequency). Therefore, in the regression for each cut-off bin, I checked whether a model with such a random slope was significantly better than a model without such a random slope. The chi-square difference between the goodness of fit between the two models was assessed (e.g., Baayen, 2008). The results from model comparisons suggested that such a random slope did not significantly improve the models, either for the high cut-off bin (p = .177) or the low cut-off bin (p = .188). Consequently, the final model for each cut-off bin allowed for only a random intercept for each participant and each test item (i.e., by-subject and by-item variability in overall reaction time were accounted for), but a by-participant random slope for phrase frequency was not included.

Finally, for the model for each cut-off bin, I checked whether the assumptions underlying linear mixed-effects regressions (e.g., Winter, 2013) were generally met. First, to investigate collinearity among the predictors, I obtained variance inflation factor (VIF) scores for the continuous predictors from R. According to Field (2009), there is no hard and fast rule about VIF scores, but one common criterion, as suggested by Myers (1990), is that a VIF score of 10 or greater should be a concern. Similarly, Loewen and Plonsky (2016) recommended that a researcher should consider removing a predictor with a VIF value of more than 10 or combining it with another predictor. In the model for the high cut-off bin, the VIF scores for the number of characters and the log frequencies for the last word, the last two words, and the last three words were 1.49, 1.97, 2.31, and 1.38 respectively. In the model for the low cut-off bin, the respective VIF scores were 1.77, 3.63, 6.11, and 2.46. The obtained VIF values were thus below the threshold level of 10.

Residual plots for the regression models for the high and low cut-off bins are shown in Figure 3 and Figure 4 respectively. Visual inspection of these plots did not indicate any obvious deviations from linearity, homoscedasticity, or normality. That is, in each model, the plot of fitted values against the standardized residuals contained randomly dispersed data points and did not suggest any obvious pattern. Moreover, the histogram and the qq-plot of residuals suggested that the residuals were normally distributed. In sum, in the regression models for the two cut-off bins, the linear regression assumptions appeared to be generally met.

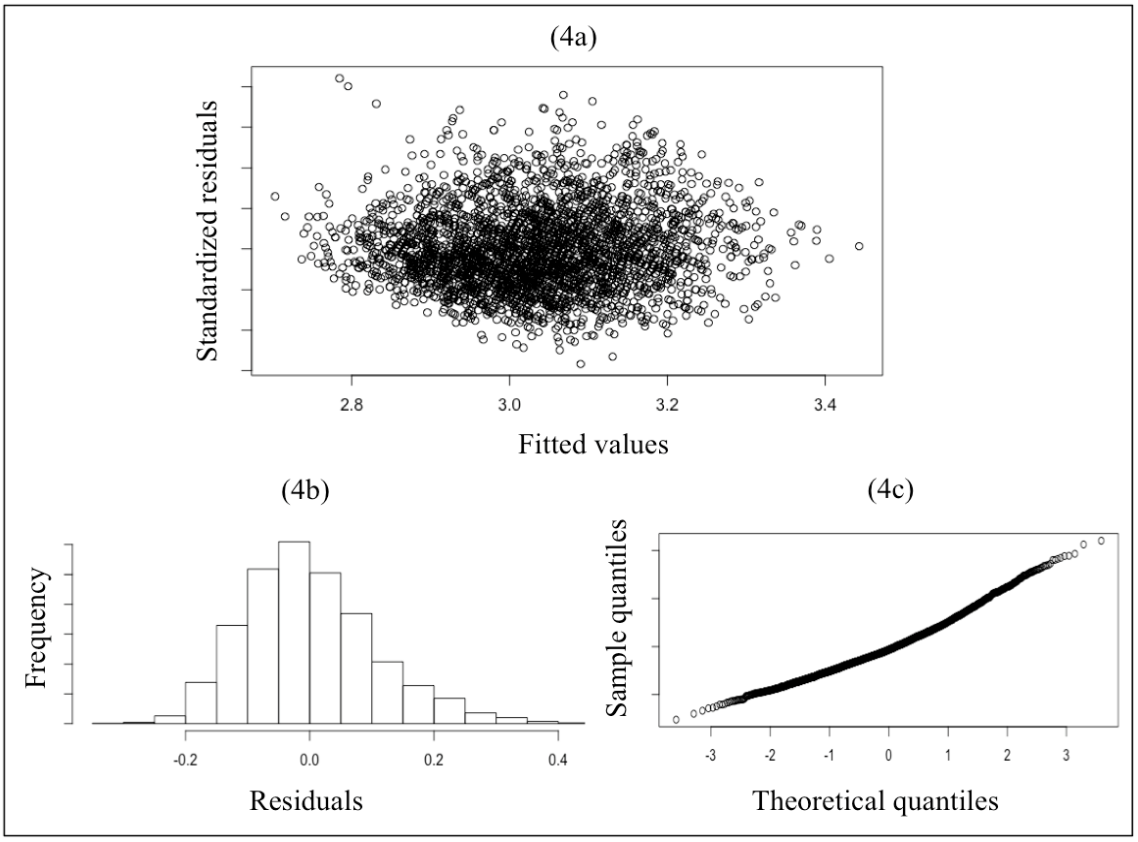

Figure 4. Residual plots for the regression model for the high cut-off bin in the phrasal acceptability judgment task: plot of fitted values against the standardized residuals (4a), residual histogram (4b), and residual qq-plot (4c)

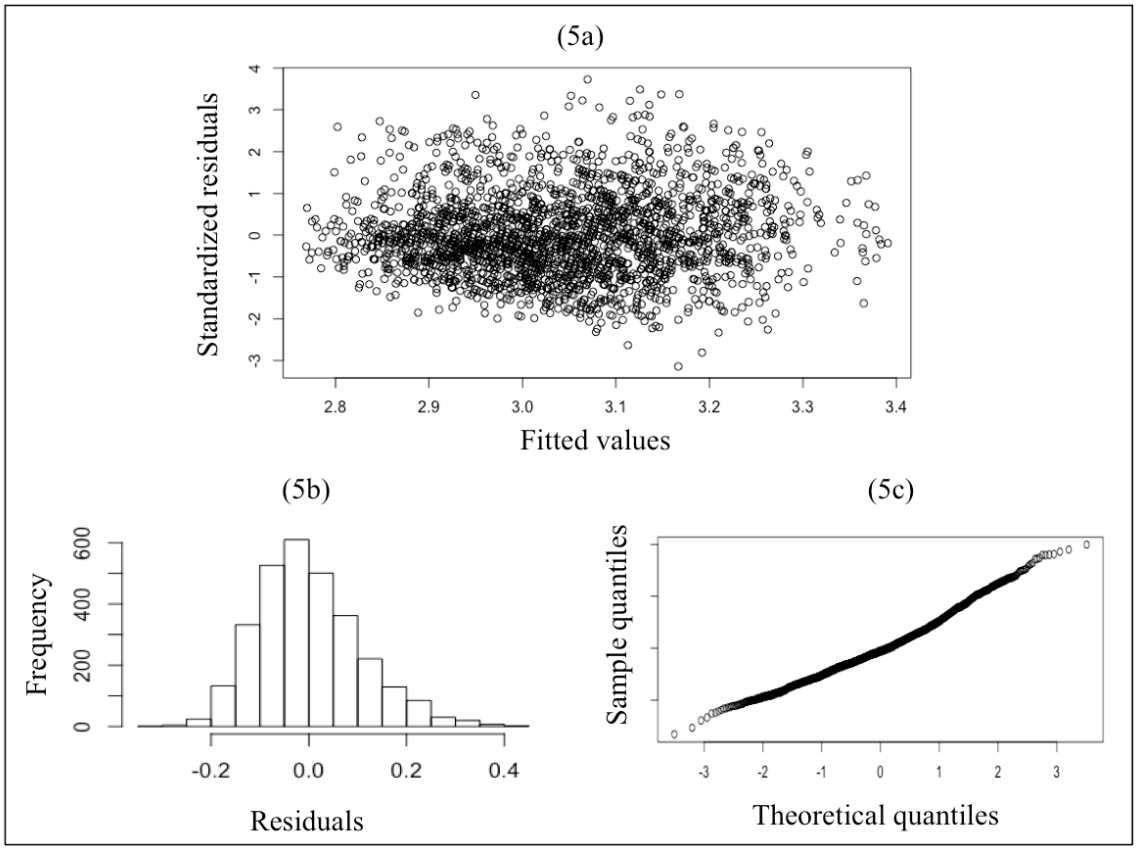

Figure 5. Residual plots for the regression model for the low cut-off bin the phrasal acceptability judgment task: plot of fitted values against standardized residuals (5a), residual histogram (5b) and residual qq-plot (5c)

2.4 Results

Table 6 shows the average reaction times from the phrasal acceptability judgment task in milliseconds (ms). The last column indicates the difference in reaction time between the high frequency condition and the low frequency condition (the baseline frequency category) in each cut-off bin in each participant group. In the high cut-off bin, in which the cut-off point for classifying a phrase as a high frequency phrase (e.g., *don't have to worry*) or a low frequency phrase (e.g., *don't have to wait*) was 12.0 occurrences per million, the native speaker mean reaction time for high frequency phrases was about 84 ms shorter than the mean reaction time for low frequency phrases. This direction of

results was also observed in the low cut-off bin, in which the cut-off point for classifying a phrase as a high frequency phrases (e.g., *don't have any money*) or a low frequency phrases (e.g., *don't have any place*) was 1 occurrence per million. That is, in this bin, the mean reaction time for high frequency phrases was approximately 48 ms shorter than the mean reaction time for low frequency phrases in the native speaker group.

Unsurprisingly, the ESL learners were generally slower than the native speakers. Overall the reaction times from the learners also had more variability, as indicated by the higher standard deviations. However, as in the case of the native speakers, the learners reacted faster to high frequency phrases than to low frequency phrases. In the high and the low cut-off bins respectively, the mean reaction times for high frequency phrases were about 116 ms and 94 ms shorter than the mean reaction times for low frequency phrases.

Table 6. Average reaction times in milliseconds from the phrasal acceptability judgment task (*SD* in parentheses)

	Phrase frequency		Frequency effects
	Low	High	(High – Low)
NS (*N*=50)			
High cut-off bin	1,029.06 (339.52)	944.94 (279.23)	-84.12
Low cut-off bin	1,003.26 (324.88)	955.48 (287.79)	-47.78
ESL (*N*=49)			
High cut-off bin	1,399.37 (556.35)	1,283.86 (476.90)	-115.51
Low cut-off bin	1,462.25 (605.05)	1,368.19 (504.81)	-94.06

Note. NS = Native English speakers, ESL = ESL learners

Table 7 reports results from the mixed-effects regressions for stimuli in the high cut-off bin. There were significant main effects of phrase frequency (χ^2 (1) = 7.13, p =

.008), participant group (χ^2 (1) = 61.69, p < .001), block order (χ^2 (1) = 115.42, p < .001), and the number of characters in the target phrases (χ^2 (1) = 21.73, p < .001). The regression coefficient (β) for each predictor indicates the change in reaction times on the base-10 logarithmic scale as a result of a one-unit change in the predictor. Because the dependent variable was on the logarithmic scale, for more meaningful result interpretations, I calculated the exponential value of each coefficient (10^{β}). The exponential function is the inverse of the logarithmic function, and in the case of binary predictors (e.g., frequency condition), the exponential value (10^{β}) expresses the average multiplicative change in reaction time between the non-reference category (e.g., high) and the reference category (e.g., low). Similarly, for continuous predictors (e.g., the number of characters), which were standardized, the exponential value expresses the average multiplicative change in reaction time associated with a one *SD* change in the predictor.[19]

As Table 7 shows, participants in both groups demonstrated sensitivity to phrase frequency. On average reaction times to high frequency phrases were 0.91 time the reaction times to low frequency phrases. That is, participants were about 9% faster when judging the acceptability of high frequency phrases when the block order, the number of characters, and substring frequency were controlled for. Moreover, the ESL learners were on average 32% slower than the native English speakers, and participants were on average 9% faster when they did the second experiment block, whether it was block A or block B, possibly due to greater task familiarity. In the experiment, the presentation of

[19] Please see Appendix F for more details about how the regression coefficients (β) based on the dependent variable on the base-10 logarithmic scale were transformed for the result interpretations.

Table 7. Mixed effects model results for the high cut-off bin in the phrasal acceptability task

Predictors	β	95% C.I.	10^{β}	*SE*	*p*
Phrase frequency (baseline = low)	-0.04	[-0.08, 0.01]	0.91	0.02	.008**
Group (baseline = NS)	0.12	[0.09, 0.16]	1.32	0.02	< .001***
Phrase frequency*group	-0.002	[-0.01, 0.02]	1.00	0.01	.841
Block order (baseline = first)	-0.04	[-0.05, -0.03]	0.91	0.004	< .001***
Number of characters	0.04	[0.03, 0.06]	1.10	0.01	< .001***
Log frequency of last word	-0.002	[-0.02, 0.02]	1.00	0.01	.840
Log frequency of last two words	-0.01	[-0.03, 0.01]	0.98	0.01	.445
Log frequency of last three words	0.01	[-0.01, 0.03]	1.02	0.01	.269

Note. R^2 marginal = .23. R^2 conditional = .54. *SE* = standard error. NS = native English speakers.

the two experimental blocks (A and B) was counterbalanced. The results suggested that there was a task familiarity effect on response latencies from block B if participants did block A first, while a task familiarity effect was present in response latencies from block A if participants did block B first. The counterbalanced design equally distributed the task familiarity effects between the two experiment blocks. However, by entering block order as an additional control variable, the portion of variance in reaction time as a result of task familiarity was accounted for statistically.

In addition, a one *SD* increase in the number of characters in the target phrases corresponded to about a 10% increase in reaction times, indicating that the participants had to spend more time reading the phrases. The interaction between phrase frequency and group was not significant (χ^2 (1) = 0.04, p = .841), and neither were the frequencies of the three subparts that differed in each target pair—that is, the last word (χ^2 (1) = 0.04, p = .840), the last two words (χ^2 (1) = 0.58, p = .445), and the last three words (χ^2 (1) = 1.22, p = .269). As in previous research using a similar model (Gyllstad & Wolter,

2016), I used the MUMIn function in R to obtain R^2 values for a mixed effects regression model. The function provides two types of R^2 values: marginal and conditional. The former is associated with the fixed effects, listed in this table, and the latter reflects the fixed and the random effects combined. In this model, the fixed effects and the random effects together can explain about 54% of the variance in the participants' reaction times.

A similar result pattern was obtained from the regression model for stimuli in the low cut-off bin. As shown in Table 8, the effects of phrase frequency (χ^2 (1) = 4.78, p = .029), participant group (χ^2 (1) = 85.83, p < .001), block order (χ^2 (1) = 125.89, p < .001), and the number of characters in the target phrases (χ^2 (1) = 19.38, p < .001) were significant. Participants in both groups were sensitive to phrase frequency. On average, they were 9% faster when judging the acceptability of high frequency phrases when the block order, the number of characters, and substring frequency were controlled for. The ESL learners were approximately 41% slower than the native English speakers. Reaction times in the second experimental block were on average 11% shorter than in the first block, and a one *SD* increase in the number of characters in the target phrases corresponded to an approximately 10% increase in reaction times. As in the high cut-off bin, the interaction between phrase frequency and group did not reach significance (χ^2 (1) = 0.15, p = .695). Also, the frequencies of the last word (χ^2 (1) = 3.75, p = .053), the last two words (χ^2 (1) = 1.32, p = .250), and the last three words (χ^2 (1) = 0.59, p = .443) were not significant predictors of response latencies.

Table 8. Mixed effects model results for the low cut-off bin in the phrasal acceptability task

Fixed effects	β	95% C.I.	10^{β}	*SE*	*p*
Phrase frequency (baseline = low)	-0.04	[-0.07, -0.01]	0.91	0.02	.029*
Group (baseline = NS)	0.15	[0.13, 0.18]	1.41	0.02	< .001***
Phrase frequency*group	-0.004	[-0.02, 0.02]	0.99	0.01	.695
Block order (baseline = first)	-0.05	[-0.06, -0.04]	0.89	0.005	< .001***
Number of characters	0.04	[0.02, 0.06]	1.10	0.008	< .001***
Log frequency of last word	0.02	[-0.01, 0.05]	1.05	0.01	.053
Log frequency of last two words	-0.02	[-0.04, 0.01]	0.95	0.02	.250
Log frequency of last three words	0.01	[-0.02, 0.04]	1.02	0.01	.443

Note. R^2 marginal = .31. R^2 conditional = .56. *SE* = standard error. NS = native English speakers

2.5 Discussion

To recap, the current experiment was conducted to test usage-based researchers' prediction that adult native English speakers and ESL learners should demonstrate frequency effects during receptive processing of compositional multi-word sequences (e.g., Bybee, 2010; Ellis, 1996, 2003, 2011, 2012; Gries & Ellis, 2015) when frequencies of the subparts of the sequences were controlled for. Such frequency effects were previously documented in adult native English speakers in the study by Arnon and Snider (2010) and the very recent study by Hernández et al. (2016). In addition, the study by Hernández et al. (2016) was the first that reported such frequency effects in English L2 learners.

The results from the current experiment were in line with the results from these two preceding studies. Both the adult native English speakers and ESL learners demonstrated frequency effects. In both stimuli cut-off bins, frequency of the target four-word phrases (high/ low) significantly predicted reaction times from both participant

groups; reaction time for high frequency phrases was significantly shorter than reaction time for low frequency phrases. This did not stem from differences in substring frequencies between the two phrases in each target pair because, in the regression analysis, substring frequencies were strictly controlled for. In addition, I controlled for other possible differences between the two phrases in each target pair which were reported as significant predictors of response latencies in phrasal acceptability tasks in previous studies (Arnon & Snider, 2010; Hernández et al., 2016). These included the block order and the number of characters in the target phrases. In short, in the current experiment, the shorter reaction time to high frequency phrases in each stimuli cut-off bin resulted from higher whole phrase frequency.

The findings from the current experiment are in line with the prediction in usage-based approaches, in which L1 acquisition is based on an accumulation of statistical information in previously-encountered L1 input. Consequently, words that co-occur more frequently in compositional phrases have stronger connections in speakers' mental linguistic representation and therefore are processed faster receptively (e.g., Bybee, 2010; Ellis, 2002, 2005, 2012; Gries & Ellis, 2015). The results do not seem compatible with an L1 representation or processing model in which the mental lexicon and abstract grammar rules are rigidly divided. In such a model, only the processing of the items in the mental lexicon, namely individual words, bound morphemes, and idioms, should demonstrate frequency effects, while the processing of compositional phrases, which are generated from abstract phrasal structure rules, are unlikely to demonstrate the effects (e.g., Prasada & Pinker, 1993; Ullman, 1999). On the other hand, the results are in line with a usage-based L1 model in which there is no complete separation of the mental

lexicon and grammar, and linguistic units of varying sizes are represented and processed based on similar general mechanisms (e.g., Abbot–Smith & Tomasello, 2006; Bybee, 2010; Ellis, 2011; Goldberg, 1995, 2006; Gries & Ellis, 2015). In such a model, the processing of morphemes, words, idioms, and compositional phrases should be frequency sensitive, and multi-word frequency must be accounted for.

Despite the possible L1–L2 differences in terms of the nature of acquisition and input (e.g., Ellis & Laporte, 1997; Muñoz, 2008), the results from the current experiment appear to support usage-based researchers' proposal that L2 acquisition may also be based on the general mechanism operating in L1 acquisition (e.g., Bybee, 2008; Ellis, 1996, 2002, 2003, 2006a, 2006b, 2008a, 2008b, 2011, 2012, 2013; Ellis & Cadierno, 2009; Ellis & Larsen–Freeman, 2009; Ellis & Wulff, 2015; Goldberg & Casenhiser, 2008; Robinson & Ellis, 2008; Römer et al., 2014). That is, as in the case of native English speakers, the human cognitive ability of chunking allows previously encountered English word sequences to be registered in ESL learners' memory and creates sequential relations between the component words. Moreover, the processing of the registered ESL word sequences in additional input later makes the strength of the relations reflect the frequency of previous encounters (Ellis, 2003, 2011, 2012). If these were not the case, the frequency effects in the English learners in the current study should not have been observed. However, given the frequency effects in the adult ESL learners, I do not argue that the input that the native English speakers and the input that the ESL learners in the current experiment had received were identical. Based on the learners' background, these learners did not start learning English in an immersion environment. Instead, they started learning English in formal English classrooms in their home country, where English is a

foreign language. A great deal of English instruction in such classrooms typically involves explicit L2 instruction (e.g., Ellis & Laporte, 1997), which can help register a novel L2 word sequence in a learner's memory before subsequent statistical fine-tuning through L2 exposure (Ellis, 2005, 2011, 2012). As discussed in the literature review, Ellis and Larsen–Freeman (2009) suggested that there may be some similarities between the input that native English speakers and adult ESL learners receive. For example, the distribution of verbs in a verb argument construction, which generally adheres to Zipf's (1935) law in L1 child-directed speech and children's speech production (Goldberg, 1999; Goldberg et al., 2004), may also be the distribution of verbs in initial verb argument constructions produced by English learners' native speaker conversation partners and ESL teachers. However, the L2 literature has also documented possible differences in the input that native English speakers and adult ESL learners have received (e.g., Ellis & Laporte, 1997; Littlemore, 2009; Muñoz, 2008). The current experiment is similar to the previous studies reporting frequency effects on ESL compositional phrase comprehension in that the frequencies of target phrases were derived from native English speaker corpora—such as COCA (Gyllstad & Wolter, 2016; Wolter & Gyllstad; 2013) and the BNC (Sonbul, 2015). More empirical support is still needed to shed light on whether phrases that are more frequent in a naturalistic English speaking environment are also more frequent in formal ESL classrooms or in the authentic English input that English learners in a foreign country are likely to receive (e.g., through English TV programs, news reports, or movies, or in their university courses). However, the frequency effects reported in the current experiment and in those ESL studies seem to support the argument that, although individuals' language experiences may differ,

frequency data from a large and adequately representative corpus probably represent the common regularities of input all speakers experience (Hoey, 2005; Wolter & Gyllstad, 2013).

As discussed, in the current experiment, the inclusion of the two stimuli cut-off bins served two purposes. First, in Arnon and Snider's (2010) study, native English speakers had significantly shorter reaction time to higher frequency phrases in both the high and the low cut-off bins. The cut-off point for classifying a phrase as a high frequency phrase (e.g., *don't have to worry*) or a low frequency phrase (e.g., *don't have to worry*) in these bins was 10 and 1 occurrences per million words respectively.[20] As noted, there have been two assumptions regarding the representation of frequently co-occurring word sequences. One holds that only a highly frequent compositional phrase—a phrase with a frequency above a threshold level—is stored as a whole and is processed faster than a phrase with a frequency below this threshold level, which is not stored as a whole but is analyzed or computed based on language grammar (e.g., Goldberg, 2003, 2006; Wray, 2002). The second proposal is that there is no such a frequency threshold. More frequent phrases are more entrenched in speakers' representation, and therefore the difference between a more frequent phrase and a less frequent phrase is quantitative, resulting from different frequencies of previous encounters. Relatively more frequent phrases should therefore be processed faster than less frequency phrases regardless of the frequency range. Based on their results, Arnon and Snider (2010) argued that native English speakers do not process only highly frequent phrases—those with frequencies

[20] These are the cut-off frequency points reported by Arnon and Snider (2010). As discussed in the method section, the cut-off frequency point for the high cut-off bin that I could derive was slightly different (12 occurrences per million words).

above the cut-off point in the high cut off bin—faster. Frequency effects were observed in both cut-off bins in a similar way, as opposed to being observed only in the high cut-off bin. That is, in the low cut-off bin, in which all phrases except one had frequency below 10 occurrences per million words, participants also processed phrase in the high frequency condition faster. Therefore, Arnon and Snider contended that their findings do not support the proposal that there is a high frequency threshold (e.g., 10 occurrences per million words) and that only compositional phrases with a frequency above this threshold level are stored a whole and are processed faster less frequent phrases. Rather, the differences in reaction times to higher and lower frequency phrases in each cut-off bin should have resulted from the relative differences in phrase frequencies. In the current experiment, therefore, I was able to replicate Arnon and Snider's (2010) findings and extended these findings to adult ESL learners.

The other reason for including stimuli from different cut-off bins in the current study was to investigate whether the ESL learners had stored sufficient accumulated statistics information (i.e., frequency of occurrences) through previous ESL exposure to demonstrate frequency effects in both cut-off bins. My initial speculation was that the ESL learners might not have had much exposure to the target phrases in the low cut-off bin (whether the high or the low frequency phrases) and thus may not demonstrate frequency effects in this bin. The results, however, suggested that the ESL learners exhibited frequency effects in both cut-off bins. The ESL learners in the current study can be characterized as being proficient enough to study in an English speaking environment. They had lived in the US for approximately 2-3 years, and given that they had to develop their English skills for US university admissions, a reasonable assumption

is that, prior to coming to the US, the ESL learners had been exposed to English input from various sources. The results in the current experiment seemed to suggest that the amount of exposure to compositional multi-word phrases that the learners had accumulated after many years of English education and after having been in the US for about 2-3 years was enough for them to exhibit sensitivity to phrase frequencies derived from the 20-million-word data combined from Fisher and Switchboard corpora, which contain natural conversations produced by native American English speakers, whether the phrases are in the low or high end of the frequency range (i.e., high or low cut-off bin).

Finally, in the current experiment, although adult L1 and L2 speakers were sensitive to phrase frequency and unsurprisingly the former group processed the target phrases significantly faster than the latter, in both cut-off bins, the regression results did not suggest a significant interaction between participant group and phrase frequency. This lack of interaction was also reported in a previous study on phrase frequency effects by Hernández et al. (2016). Interestingly, the absence of the interaction differed from the consistent finding in empirical research on single word recognition (e.g., Diependaele et al., 2013; Duyck et al., 2008; Whitford & Titone, 2012), according to which frequency effects were observed from both L1 and L2 speakers in recognition tasks but were stronger in L2 learners as indicated by a significant interaction between word frequency and participant group. Such a finding from single word recognition studies has prompted researchers to propose *the lexical entrenchment hypothesis* (e.g., Diependaele et al., 2013). The important idea is that due to lower English proficiency and less English input, English words are less well entrenched in ESL learners' mental representation than

in native English speakers' representation. Therefore, processing of ESL words requires more effort generally, but particularly greater effort is required when ESL words have low frequency. As result, the difference in processing high and low frequency English words in L2 speakers is more pronounced than in L1 speakers.

In Hernández et al.'s (2016) study, in addition to the phrasal acceptability task, the researchers conducted a lexical decision task in which the same participants as in the phrasal acceptability task judged whether strings of letters presented on a computer screen were English words. The stimuli included 20 high frequency words and 20 low frequency words. The words in these categories had mean frequencies of 109.3 (*SD* = 133.99) and 5.45 (*SD* = 2.85) occurrences respectively based on CELEX database (Baayen, Piepenbrock, & Gulikers, 1995), which contains frequency data from the 17.9-million-word COBUILD Corpus from the University of Birmingham. Based on this lexical decision task, as in previous single word recognition research, Hernández et al. (2006) reported frequency effects in both native speakers and English learners, but stronger effects in the learner group, as indicated by a significant frequency by group interaction. Comparing these results against those from the phrasal acceptability task, Hernández et al. (2006) speculated that stronger frequency effects in the ESL learners in the phrasal acceptability task may also exist but may not have been observed because the mean frequency differences between high and low frequency phrases (16.33 and 2.76 occurrences per million words in the high and low cut-off bins respectively) were too low compared to the mean frequency differences between high and low frequency words in the lexical decision task (103.85 occurrences per million words). In the current experiment, like Hernández et al. (2006), I used a part of the stimuli from Arnon and

Snider (2010) and similarly observed no interaction between participant group and frequency. Based on the frequency data I was able to derive, in the current experiment, the mean frequency differences between high and low frequency phrases in the high and the low cut-off bins were 20.13 and 4.41 occurrences per million words respectively. These were thus similar to the mean frequency differences between the high and low frequency phrases in the two respective bins reported in Hernández et al.'s (2006) study. Consequently, in line with Hernández et al.'s (2006) observation, I believe that future comprehension studies could further investigate whether there are stronger frequency effects for multiword sequences in L2 speakers compared to in L1 speakers (as indicated by a group and frequency interaction) when stimuli with a wider range of frequencies are used. Such future studies are particularly interesting given the rather limited existing ESL research on compositional multi-word sequence processing with substring frequency control. The results may indicate similarities or differences between single word and compositional phrase receptive processing in L1 and L2 speakers.

CHAPTER 3: EXPERIMENT II

3.1 Research question and prediction

Conducted to answer the second research question, this experiment tested frequency effects on compositional multi-word sequence production when sub-part frequencies of the target sequences were controlled for. The same stimuli from the first experiment were used with the same participants, so that the results illuminated whether these same participants demonstrated frequency effects in both language comprehension and production. While frequency effects had been documented in spontaneous speech production (Arnon & Priva, 2013, 2014), I used an elicited oral production task because previous studies based on this task also demonstrated phrase frequency effects (Arnon & Priva, 2013; Bannard & Matthew, 2008) and because, with this type of task, I was able to elicit production of all the target phrases in the first experiment from the same participants. When compared to results from the first experiment, the results from this experiment should therefore suggest if the different nature of tasks influenced frequency effects.

As discussed in the first chapter, two previous lab-based studies investigated frequency effects on compositional multi-word phrase productive processing with substring frequency control, that is, Bannard and Matthew (2008) and Arnon and Priva (2013). Following these studies, in the current experiment, I operationalized language production as the phonetic durations of the first three words (e.g., *don't have to*) in the target phrases (e.g., *don't have to worry*) in a speech elicitation task. The part of this experiment with the native English speakers was a partial replication of Arnon and Priva's (2013) speech elicitation experiment. An assumption was that the first three

words in a more frequent phrase (e.g., *don't have to worry*) should be produced faster than the first three words in a less frequent phrase (e.g., *don't have to wait*) due to the stronger relations between words (e.g., Bybee, 2010). Based on the previous two studies, the adult native English speakers in this experiment should demonstrate frequency effects, while it remained to be seen whether similar results would be observed from the ESL learners because prior to the current experiment they had been no relevant L2 production research with substring frequency control.

3.2 Method

3.2.1 Participants

The same participants in the first experiment completed the present experiment. The purpose was to control for individual differences across the two tasks. Because the target phrases in the first experiment were also used in this experiment, the participants completed this experiment at least two days after the first. This was to reduce the possibility that the processing of the target phrases in the first experiment would affect the processing in the current experiment. Moreover, all participants completed the first before the second experiment so that their production of the target phrases in the second experiment did not influence their acceptability judgments in the first experiment.

3.2.2 Materials

The same 28 target pairs from the first experiment were divided into two blocks: C and D. As in the first experiment, one variant from each target pair was randomly assigned to each block, and each block consisted of 14 high frequency variants and 14 low frequency variants from the target pairs. Each block therefore contained 28 target

phrases. In addition, in each block, there were 28 fillers which were possible English phrases. Thus, in each block, participants saw 56 phrases. The fillers in this experiment consisted of (1) grammatical fillers in the first experiment (e.g., *hold a green bag*) and (2) grammatical counterparts (e.g., *the girl didn't sleep*) of ungrammatical fillers in the first experiment (e.g., *girl the didn't sleep*). The purpose was to make the two tasks as comparable as possible in terms of participants' lexical exposure. As in the first experiment, lexical overlap between the target phrases and the fillers, listed in Appendix G, was also minimized.

3.2.3 Procedure

I ran this experiment on PsychoPy, a free software program for psychology and linguistics experiments (Peirce, 2007), downloadable from http://www.psychopy.org/. The participants sat in a quiet room and completed a phrase elicitation task in front of a computer screen. Each participant completed both experiment blocks. The procedure in this experiment is shown in Figure 6.

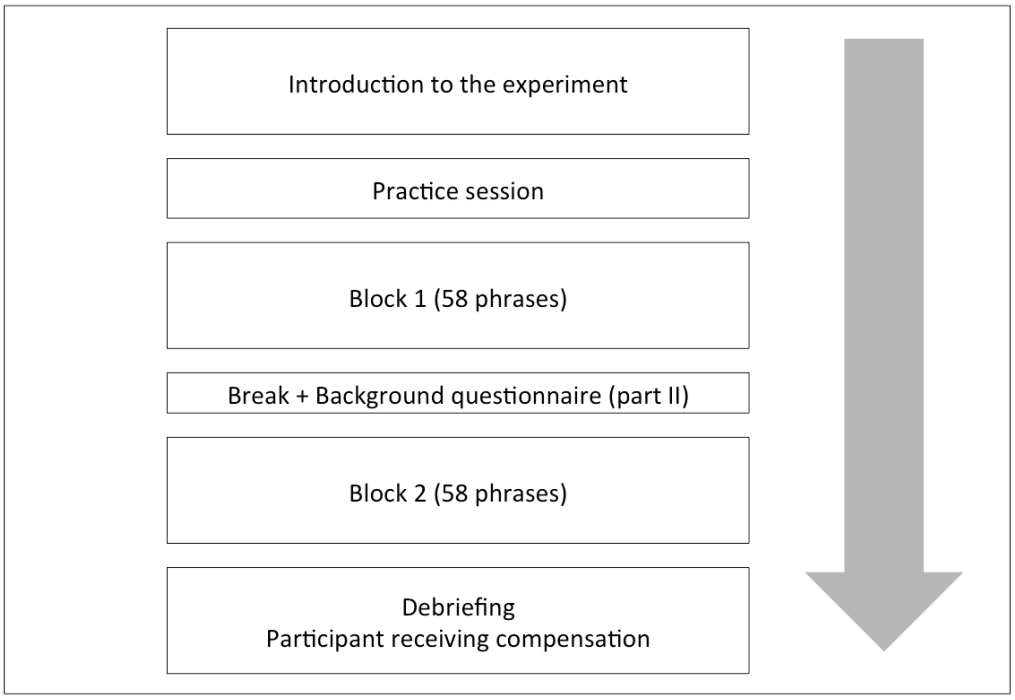

Figure 6. Procedure in the elicited production task

Because the goal of the present experiment was to investigate the effects of the frequency of the whole target phrase on speech production durations, the instruction told the participants to read the phrase as soon as the phrase disappeared (i.e., after seeing the whole phrase). In line with previous research using an elicited production task to investigate frequency effects (e.g., Ellis et al., 2008; Janssen & Barber, 2012; Tremblay & Tucker, 2011), I instructed the participants to read the phrase as fast and as accurately as they could. The exact instruction is shown in Figure 7.

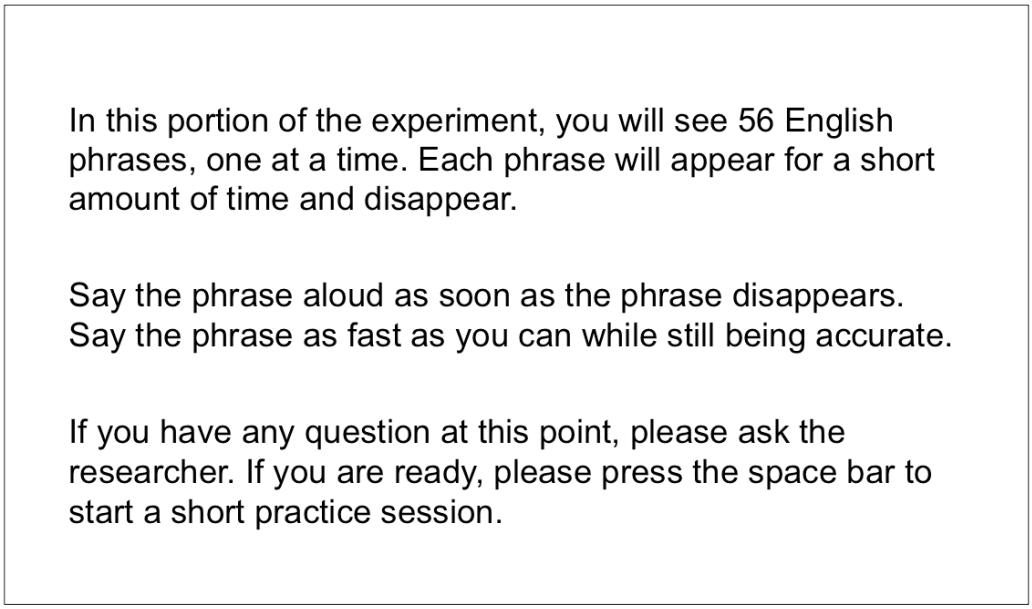

Figure 7. Instruction in the elicited production task

In this experiment, participants saw the target phrases on the screen one at a time and in their entirety (font: Arial; size: 36; position: center) for a fixed amount of time (1,700 milliseconds). The phrases were in the lower case, except for the first person personal pronoun and proper nouns. Based on a pilot study, the interval between the end of a phrase presentation and the time the next phrase appeared was set at 2,500 milliseconds because this duration was found to be long enough for speech production of both participant groups. There were also six practice items at the beginning of each experiment block. Before the practice section began, I additionally instructed the participants that they should continue the experiment even when they felt they made a mistake (e.g., said a word wrong). The participants were also informed that later during the experiment, they did not have to press any computer key because the program would run automatically and that their voice would be recorded. The program started recording participants' production duration from the moment each phrase disappeared from the

computer screen to the moment the following phrase appeared. Thus, the recorded production duration for each phrase was 2,500 milliseconds, including any silence during this period.

Unlike Arnon and Priva (2013), I used a within-subject counterbalanced design to control for participants' individual variability in processing phrases in the two blocks. As noted, Arnon and Priva (2013) addressed a possible repetition effect resulting from a participant's reproduction of the same trigrams from a target pair (e.g., *don't have to worry* and *don't have to wait*) by using a between-subject design—that is, each participant read only one variant from each pair in their elicited oral production experiment (i.e., one participant's production of *don't have to* in *don't have to worry* was compared against another participant's production of this trigram in *don't have to wait*). The researchers entered the average production durations of each participant (across all target stimuli) in their regression model as a predictor of production durations to account for individual variability. However, because the repetition effect was also possible in the first experiment, and I addressed it by using a within-subject design and two counterbalanced blocks with an intervening break, I also used a within-subject counterbalanced design in the current experiment to both maintain consistency and to control individual variability. Therefore, the participants in each group (native English speakers and ESL learners) were randomly and equally divided to complete either block C or block D first. Each participant completed both blocks, which were separated by a break in which they filled out the second part of the background questionnaire. As in the first experiment, the break was included to further reduce a repetition effect resulting from the participant's production of the same first three words from each target pair.

Across the two blocks, each participant thus saw the two variants from each target pair. The phrases in each block were presented in a random order. The whole experiment took approximately 25 minutes. Participants received 20 USD after the completion of this experiment.

3.3 Analysis

For each phrase, PsychoPy recorded the 2,500-millisecond interval in a separate short audio file and put the file in a folder created for each block per participant. That is, for each participant, the program created two folders, one for block C and the other for block D. Each of the two folders consisted of 56 separate short audio files for the 28 target phrases and the 28 fillers in the block. Because of the stimuli randomization, the order and the name of the short audio files in each folder differed. Therefore, to prepare the data for the statistical analysis, I did the following.[21]

First, I concatenated the short audio files in each of the two folders from each participant. The purpose was to create only one long audio file per block per participant. To do so, I used two computer software programs: (1) R and (2) PRAAT, a free computer software package for speech and phonetic analyses (Boersma & Weenink, 2010), available at http://www.fon.hum.uva.nl/praat/. I first created and ran a script in R which in turn created two text files: a list of the 28 target phrases in a block in the alphabetical order and a list of audio file names that corresponded to the order. These two text files were for the concatenation of the short audio files in each block from each participant. A screen shots of this R script is in Figure 8. By running this R script, therefore, I excluded

[21] I am grateful to Karthik Durvasula and Qian Luo for their help and suggestions.

production of fillers from the analysis. This was possible because when this experiment was created in PsychoPy, I entered a code identifying whether a phrase was a target or a filler.

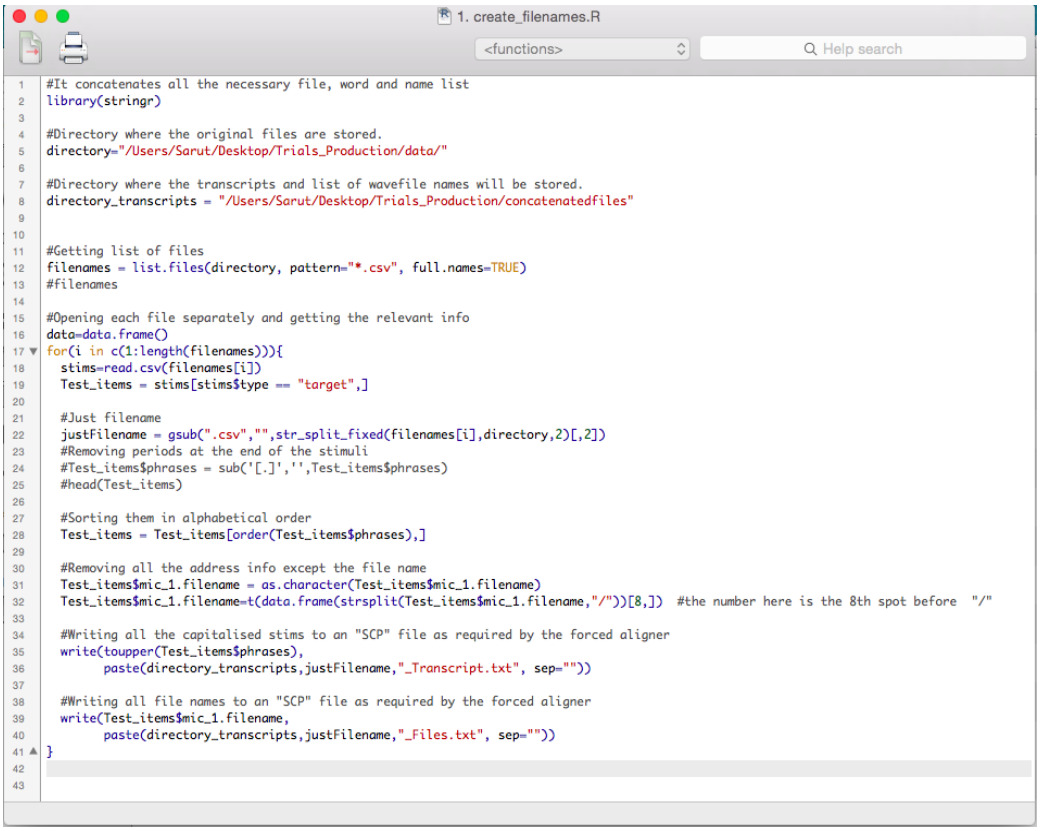

Figure 8. Screen shot of the R script used for audio file concatenation

Next, I created and ran a PRAAT script which automatically concatenated the 28 audio files in each folder from each participant using the two text files created in the preceding step. A screen shot of this PRAAT script is in Figure 9. The resulting product was one long audio file per block per participant. Among the 103 participants in the first experiment, one native speaker and one ESL learners did not return to the lab to do the second experiment; therefore, 101 participants (50 native speakers and 51 ESL learners) completed this experiment. Because there were two experiment blocks, 202 long audio files were created. Each audio file contained the production of the 28 target phrases in

the block in the same (alphabetical) order, together with any silence preceding or following the production of each phrase within the 2,500-millisecond span.

2.soundfile_concatenation_new

```
#Concatenates sound files from a list

clearinfo

#Directory where the list of filenames is stored - files created in R
directory_list$ = "/Users/Sarut/Desktop/Trials_Production/concatenatedfiles/"

#Directory where the actual data is stored
data_directory$ = "/Users/Sarut/Desktop/Trials_Production/data/"


#Getting names of all the textfiles with the soundfile names
Create Strings as file list: "fileList", directory_list$+"*_Files.txt"

#Getting number of files
numFiles = Get number of strings


for fileNum from 1 to numFiles

	#Getting filename
	selectObject: "Strings fileList"
	fileName$ = Get string: fileNum
	#print 'fileName$' 'newline$'

	#getting only filename without extension
	onlyFileName$ = fileName$ - "_Files.txt"
	#print 'onlyFileName$' 'newline$'

	#Soundfile foldername
	soundFolderName$ = onlyFileName$+"_wav"
	print 'soundFolderName$'

	#Reading the file list
	Read Strings from raw text file... 'directory_list$''fileName$'
	num_of_strings=Get number of strings
	#print 'num_of_strings' 'newline$'

	#getting Stringobject Name
	soundStringObjectName$ = onlyFileName$+"_Files"
	#print 'soundStringObjectName$'

	#iterating thru each soundfile name to concatenate
	for i from 1 to num_of_strings
		select Strings 'soundStringObjectName$'
		string$ = Get string... i

		#reading the soundfile
		Read from file... 'data_directory$''soundFolderName$'/'string$'
		Rename... soundfile

		#either start a new file, or concatenate
		if (i = 1)
			select Sound soundfile
```

Figure 9. Screen shot of the PRAAT script used for audio file concatenation

The next step was to identify the production durations of the target segments (i.e., the first three words) in the target phrases in each long audio file. First, I time-aligned each audio file, using FAVE–Extract (Rosenfelder et al., 2014) and Prosodylab–Aligner (Gorman, Howell, & Wagner, 2011), available through the Dartmouth Linguistic Automation (DARLA) web interface (Reddy & Stanford 2015). To do so, I uploaded each long audio file and its transcript to http://darla.dartmouth.edu/uploadtxttrans. The resulting product was a text grid in which each word in the target phrases was aligned

with its duration. I subsequently opened each text grid and the corresponding audio file in PRAAT to check whether the identification of the beginning and the end of the production of each word was accurate. An example is shown in Figure 10. The top row is the sound wave from the production of the phrase *don't have to worry* in one of the long audio files, while the bottom row shows the relevant part of the corresponding text grid obtained from DARLA. The number (0.102 seconds) under the highlighted word (*don't*) was the production duration of the word identified by DARLA. Because there were 202 long audio files, 202 DARLA text grids were created. The automatic time alignment was time-saving because I did not have to open each long audio file in PRAAT and type the corresponding transcription and identify production durations manually.

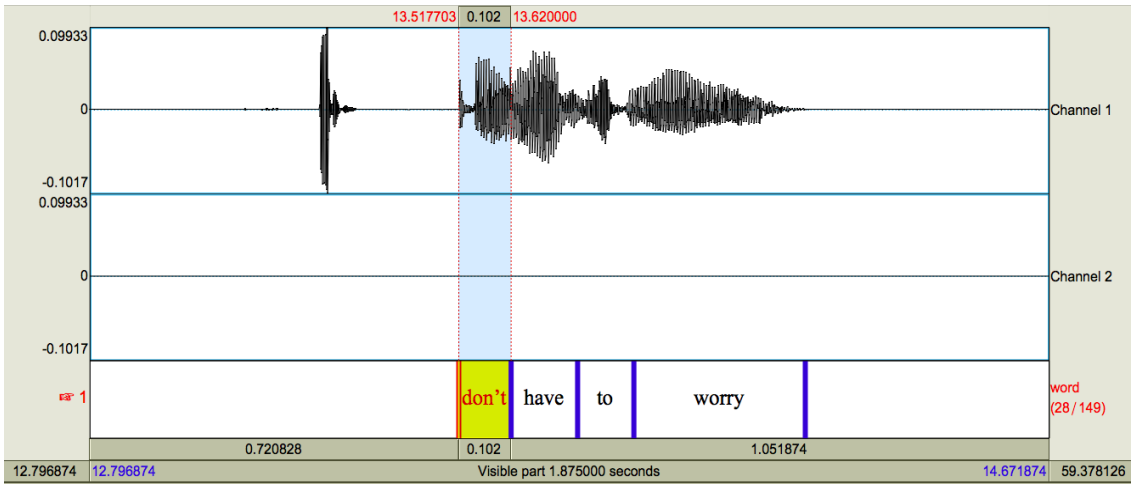

Figure 10. Example of the text grids obtained from DARLA and the corresponding audio file when opened in PRAAT

However, in some cases and especially in the text grids based on the speech production of the ESL learners, I found that the identification of individual word durations based on the text grids from DARLA was not always accurate. As a result, while opening each textgrid from DARLA with the corresponding audio file in PRAAT, I made necessary manual adjustments to each text grid to achieve greater accuracy.

Because adjusting the production duration of every word in the target segment (e.g., *don't*, *have*, and *to*) would have been extremely time consuming, to obtain a final textgrid, I combined the production durations of the three words in a target segment in each phrase, so that I could focus only on the accuracy of the identification of the beginning and the end of the target segment (e.g., *don't have to* in *don't have to worry*). For each text grid, I checked for accuracy twice to obtain the final text grid. An example of the final text grids after my manual adjustments when opened with the corresponding audio file in PRAAT is in Figure 11, in which the number (0.321 seconds) under the highlighted target segment (*don't have to*) was the production duration of the segment.

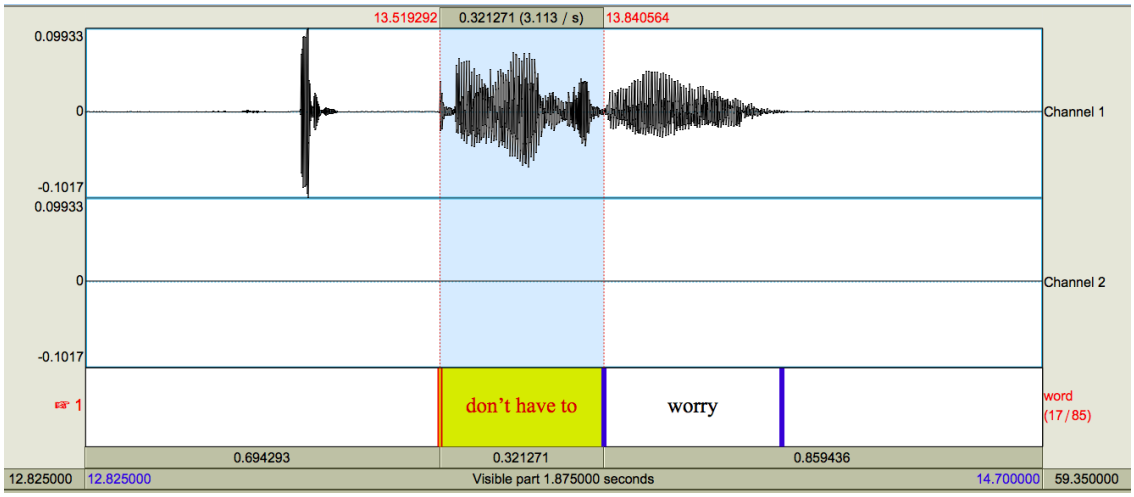

Figure 11. Example of final text grids and the corresponding audio file when opened in PRAAT

Subsequently, I created and ran another PRAAT script which converted the final text grids into data in an Excel format for the statistical analysis. This PRAAT script is shown in Figure 12. Information about the participant group, block order, phrase frequency condition (low, high), the cut-off bin (low, high), the number of syllables of the target segments and the whole phrases, and the frequencies of the subparts of the

target phrases were entered into the excel file at this stage.

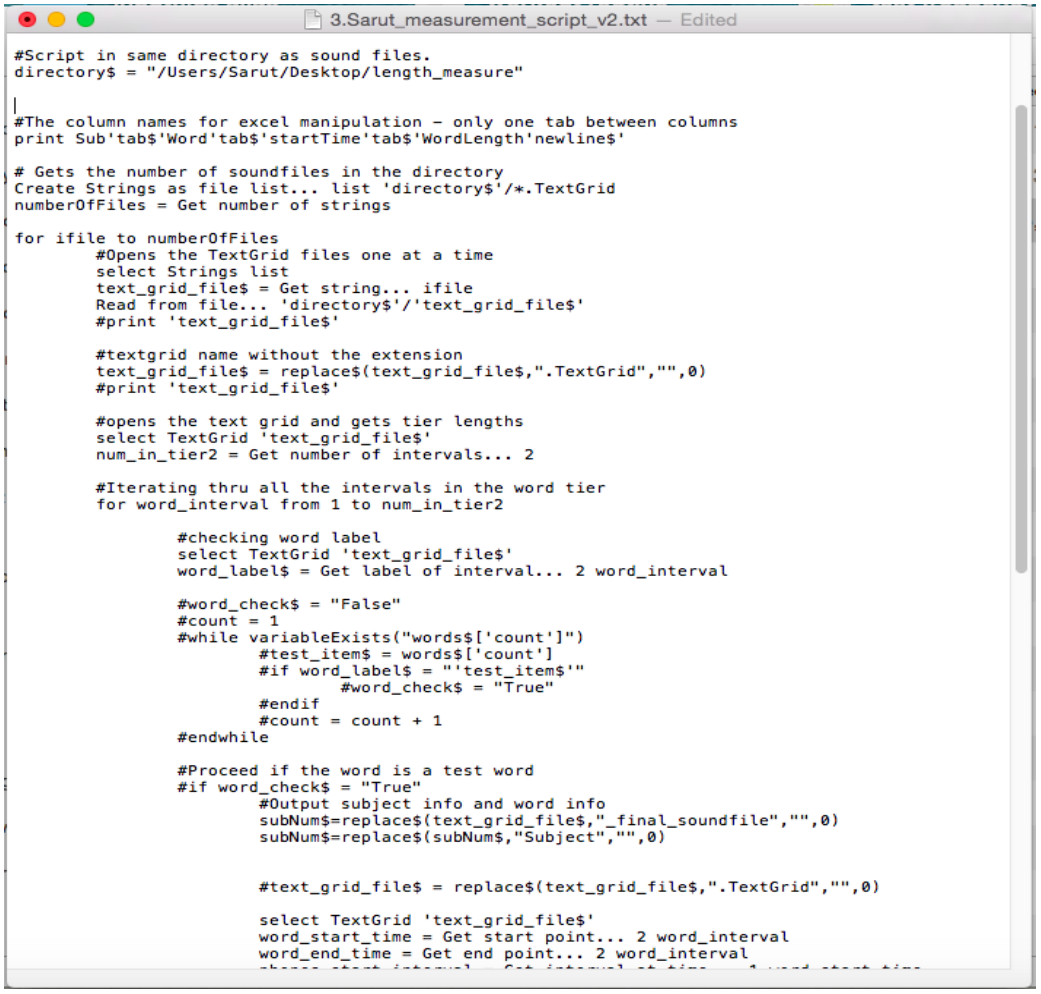

Figure 12. Screen shot of PRAAT script used to convert final time-aligned text grids to data in the Excel format

As noted, in the analysis for this experiment, I had to exclude one native English speaker and one ESL learner because they did not return to the lab to do the current experiment. Moreover, I excluded another ESL learner who did not strictly follow the directions. That is, unlike the other participants, who produced the target phrases as fast as they could, this participant occasionally slowed down and deliberately emphasized words in the target phrases. I also removed another ESL learner as an outlier because this participant spoke very slowly and thus the mean production duration from this participant

exceeded +2 *SD*s from the ESL learners' group mean in several conditions. The remaining participants consisted of 50 native English speakers (Male = 18, Female = 32) and 49 ESL learners (Male = 17, Female = 32).

The mean production accuracy in the native speaker group and the ESL learner group was 99% (Min = 96%, Max = 100%, *SD* = 0.01) and 97% (Min = 91%, Max = 100%, *SD* = 0.03) respectively. I subsequently removed inaccurate production, which consisted of incorrect and incomplete responses, from the analysis. Incorrect responses were instances in which participants said a word wrong. Incomplete responses included responses in which participants started saying a phrase before it disappeared from the computer screen. Because PsychoPy started recording a response at the moment each phrase disappeared, if participants started saying a phrase before that point, a part of the production was not recorded, and the production for that phrase was thus incomplete. As discussed, the purpose of the current experiment was to investigate phrase frequency effects, and in line with the previous relevant research (Arnon & Priva, 2013; Bannard & Matthew, 2008) participants should be exposed to the whole phrase before they started producing the target phrases. Moreover, incomplete responses included instances in which participants omitted a word or stopped (e.g., coughed) during production of a phrase. No participant was unable to produce the target segment within the 2,500-millisecond time limit. Finally, in each participant group, production durations that fell outside +/- 2 *SD*s from the group mean in each frequency condition in each cut-off bin were removed. This resulted in an exclusion of about 4% of the correct responses from each participant group.

As in the first experiment, for the analysis, I ran a mixed-effects regression model in R separately for each cut-off bin. Table 9 lists all the predictors of the dependent variable, that is, the duration of the first three words in the target pairs (e.g., *don't have to worry* and *don't have to wait*).

Table 9. Explanatory variables in the second experiment

Variable	Type	Level
Phrase frequency condition	Categorical	low*, high
Group	Categorical	NS*, ESL
Block order	Categorical	first*, second
Number of syllables in the target segment	Continuous	
Log frequency of last word	Continuous	
Log frequency of last two words	Continuous	
Log frequency of last three words	Continuous	

Note. NS = native English speakers; ESL = ESL learners. An asterisk marks the reference category for each categorical variable.

As in the first experiment, frequency condition of the target phrases and participant group were the main predictors of interest, and I investigated the interaction between these two variables. The other explanatory variables were control variables, entered to account for other differences between the two phrases in each pair that may have affected participants' production durations. The block order was included because participants may produce the target segments faster in the second experiment block, whether it was block C or D, due to task familiarity. My additional expectation was that the participants' production durations could increase as a function of the number of syllables in the target segment. Moreover, substring frequencies were entered as control variables, and I did a log transformation of the dependent variable and the frequencies of these subparts to reduce skewness of the data. I also assigned a reference category, marked with an asterisk in Table 9, for each categorical variable. Furthermore, to

illustrate the magnitude of the explanatory variables, I standardized the continuous predictors. The coefficients of standardized numerical predictors therefore indicated the change in production durations on the logarithmic scale as a result of a one standard deviation change in the continuous predictors.

Next, for the model for the high cut-off bin, I checked whether the assumptions underlying linear mixed-effects regressions were met. First, to investigate collinearity among continuous predictors, I obtained VIF scores from R. The VIF score for the number of syllables in the target segment was 1.12. Log frequencies of the three subparts—the last word, the last two words, and the last three words—had respective VIF scores of 1.82, 2.13, and 1.33. Therefore, the obtained VIF scores were below the threshold level of 10 (e.g., Field, 2009; Loewen & Plonsky, 2016). Residual plots for the regression model for the high cut-off bin are shown in Figure 13. No obvious deviations from linearity, homoscedasticity, or normality were detected. That is, the plot of fitted values against the standardized residuals seemed to contain randomly dispersed data points and did not reveal any obvious pattern, and the residual histogram and qq-plot suggested that the residuals were normally distributed. In sum, in the model for phrase production in the high cut-off bin, the linear regression assumptions were generally met.

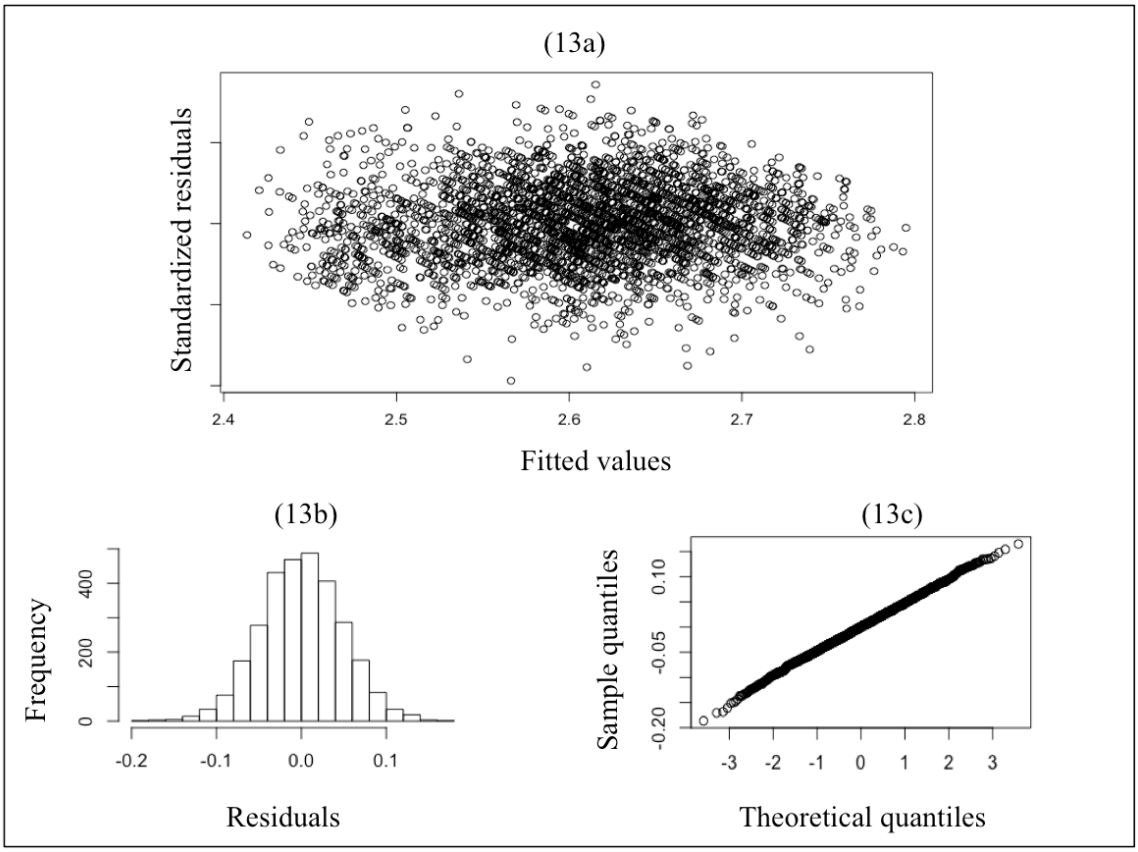

Figure 13. Residual plots for the regression model for the high cut-off bin in the production task: plot of fitted values against the standardized residuals (13a), residual histogram (13b), and residual qq-plot (13c)

In the model for the low cut-off bin, however, log frequencies of last word and the last two words of the target phrases had high VIF scores of 10.81 and 11.22 respectively. Because the values exceeded 10, at least one of these variables should be removed or combined with another predictor (Loewen & Plonsky, 2016). The two VIF values were very similar, and I found that removing either of them reduced the VIFs of all the remaining continuous predictors to below 10. I therefore needed to find a basis for the removal of one of these two control variables. In the regression model for phrase production for the high cut-off bin, which will be described in the following result section, I found that log frequency of the last two words was a significant predictor of

participants' production durations. As a result, for the low cut-off bin, I decided to keep log frequency of the last two words and removed log frequency of the last word. After the removal, the VIF values for the number of syllables of the target segment, frequency of the last two words, and frequency of the last three words were 1.02, 2.01, and 2.01, respectively. None of these were above 10. Moreover, as illustrated in Figure 14, in the resulting model, the data points in the plot of fitted values against the standardized residuals appeared randomly dispersed and did not reveal any obvious pattern, so there was no obvious deviation from linearity or homoscedasticity. Also, the residual histogram and qq-plot suggested that the residuals from the model were normally distributed.

Finally, as in the first experiment, for the model for each cut-off bin, I compared models to determine whether adding a by-participant random slope for frequency helped improve the model, and found that such a random slope did not significantly improve the models, either in the high cut-off bin (p = .808) or the low cut-off bin (p = .964). The random slope was consequently not included.

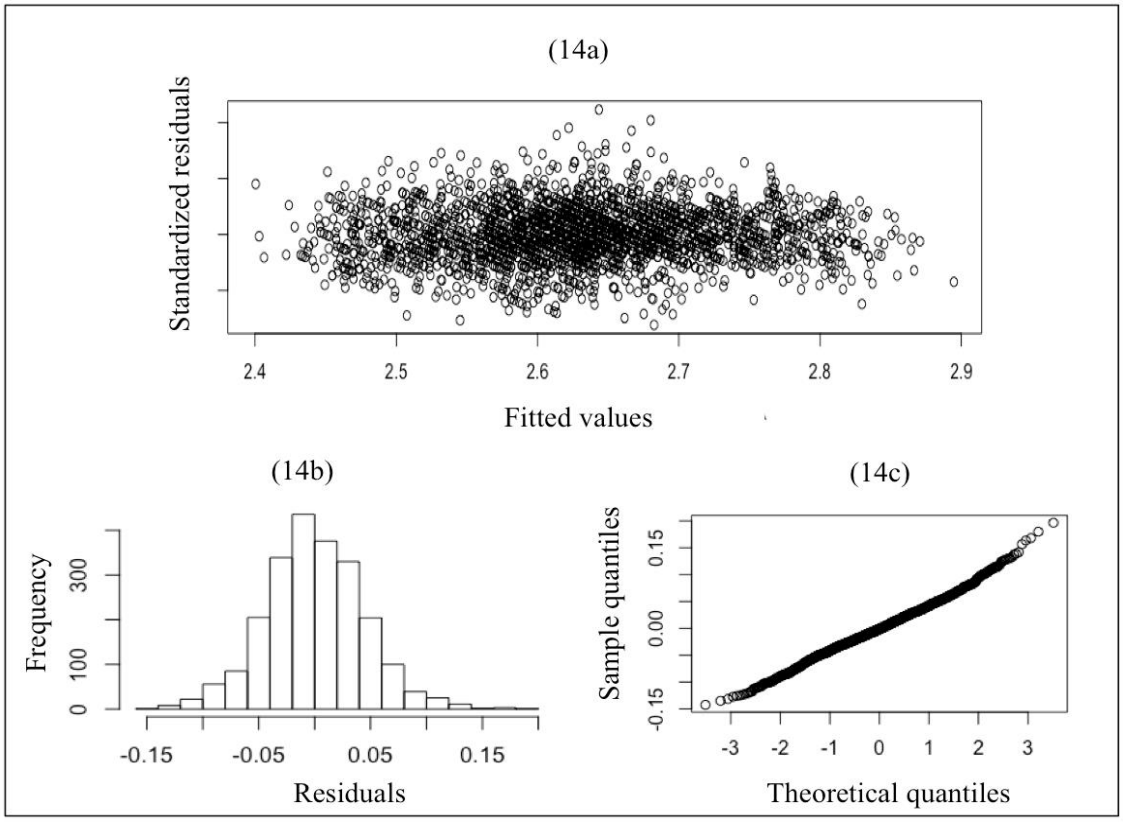

Figure 14. Residual plots for the regression model for the high cut-off bin in the production task: plot of fitted values against the standardized residuals (14a), residual histogram (14b), and residual qq-plot (14c)

3.4 Results

The average production duration for the target segments in ms in each condition is reported in Table 10. In the native speaker group, in both high and low cut-off bins, production durations for the target segments (e.g., *don't have to*) inside high frequency phrases (e.g., *don't have to worry*) were shorter than production durations for the target segments inside low frequency phrases. That is, in the two respective bins, the mean production durations for the target segments in high frequency phrases were about 18 ms and 13 ms shorter than the mean production duration for the target segments in low frequency phrases.

Although the ESL learners were generally slower than the native speakers, the results from the learners were in the same direction: the learners also produced the target segments in high frequency phrases faster. In the high and the low cut-off bins, the mean production durations for the target segments in high frequency phrases were approximately 18 and 17 ms shorter than the mean production durations for target segments in low frequency phrases, respectively.

Table 10. Average production durations of the target segments in milliseconds from the elicited production task (*SD* in parentheses)

	Phrase frequency		Frequency effects
	Low	High	(High – Low)
NS (*N*=50)			
High cut-off bin	397.23 (79.05)	379.15 (75.74)	-18.08
Low cut-off bin	411.87 (90.44)	399.26 (88.04)	-12.61
ESL (*N*=49)			
High cut-off bin	450.23 (75.99)	432.23 (74.23)	-18.00
Low cut-off bin	483.67 (107.91)	466.96 (93.44)	-16.71

Note. NS = Native English speakers, ESL = ESL learners

However, results from the mixed effects regression models for either cut-off bins suggested that phrase frequency did not significantly predict production durations of the target segments. Table 11 reports regression results from the high cut-off bin. The dependent variable was the production durations of the target segments on the base-10 logarithmic scale. As in the first experiment, for more meaningful result interpretations, I calculated the exponential value (10^{β}) for each regression coefficient (β). Although the coefficient for phrase frequency was negative and was thus in the expected direction, the effects of phrase frequency did not reach significance (χ^2 (1) = 2.35, p = .125) when the other differences between the two phrases in the target pairs were controlled for.

Table 11. Mixed effects model results for the high cut-off bin in the elicited production task

Fixed effects	β	95% C.I.	10^{β}	*SE*	*p*
Phrase frequency (baseline = low)	-0.02	[-0.05, 0.01]	0.95	0.01	.125
Group (baseline = NS)	0.06	[0.03, 0.07]	1.15	0.01	< .001***
Phrase frequency*group	0.002	[-0.004, 0.01]	1.00	0.004	.494
Block order (baseline = first)	-0.01	[-0.011, -0.005]	0.98	0.002	< .001***
Number of syllables in target segment	0.02	[0.01, 0.03]	1.05	0.06	< .001***
Log frequency of last word	0.004	[-0.01, 0.02]	1.01	0.01	.595
Log frequency of last two words	-0.02	[-0.03, -0.01]	0.95	0.01	.010**
Log frequency of last three words	0.01	[-0.01, 0.02]	1.02	0.01	.270

Note. R^2 marginal = .23. R^2 conditional = .70. *SE* = standard error. NS = native English speakers

However, there were significant main effects of participant group (χ^2 (1) = 27.94, p < .001), block order (χ^2 (1) = 16.75, p < .001), and the number of syllables in the target segments (χ^2 (1) = 12.78, p = < .001). That is, the ESL learners were on average 15% slower than the native English speakers. Participants were on average 2% faster when they did the second experiment block, whether it was block C or block D. As in the phrasal acceptability task, this could have resulted from greater task familiarity. Also, a one *SD* increase in the number of syllables in the target segments led to about a 5% increase in production durations. This was thus in the expected direction. The interaction between phrase frequency and group (χ^2 (1) = 0.47, p = .494) was not significant, and neither were the frequencies of the last word (χ^2 (1) = 0.28, p = .595) and the last three words (χ^2 (1) = 1.22, p = .270). However, the effects of frequency of the last two words in the target segment (χ^2 (1) = 6.60, p = .010) were significant. The negative coefficient for this predictor suggested that a one *SD* increase in the log frequency of the last two words in the target phrases (e.g., *to worry* in *don't have to*

worry) corresponded to a 5% decrease in the production durations of the target segment (e.g., *don't have to).* In light of usage-based approaches (e.g., Bybee, 2010), this could mean that the higher frequency, and thus the stronger relation, between the last word in a target segment (e.g., *to*) and the last word in a phrase (e.g., *worry*) made the participants produce the last word in the target segment faster, leading to shorter production durations of the target segment (i.e., the first three words in the target phrases). Arguably, this could be interpreted as evidence for frequency effects, although in this case the effects were based on frequencies of the last two words in the target phrases, not whole phrase frequencies.

Table 12 reports results from the regression model for phrase production in the low cut-off bin. As in the model for the high cut-off bin, the coefficient for phrase frequency was negative and was thus in the expected direction, but phrase frequency effects did not reach significance (χ^2 (1) = 3.21, p = .073) when the other differences between the two phrases in the target pairs were controlled for. There were however significant main effects of participant group (χ^2 (1) = 35.95, p < .001), block order (χ^2 (1) = 12.42, p < .001), and the number of syllables in the target segments (χ^2 (1) = 122.86, p < .001). That is, compared to the native speakers, the ESL learners were on average 17% slower when they produced the target segments. Moreover, participants were on average 2% faster when they did the second experiment block, and a one *SD* increase in the number of syllables in the target segments led to about a 15% increase in production durations. The interaction between phrase frequency and group (χ^2 (1) = 0.39, p = .529) was not significant, and neither were the frequencies of the last two words (χ^2 (1) = 0.23, p = .635) and the last three words (χ^2 (1) = 0.97, p = .324).

Table 12. Mixed effects model results for the low cut-off bin in the elicited production task

Fixed effects	β	95% C.I.	10^{β}	*SE*	*p*
Phrase frequency (baseline = low)	-0.03	[-0.06, -0.01]	0.93	0.02	.073
Group (baseline = NS)	0.07	[0.05, 0.09]	1.17	0.01	< .001***
Phrase frequency*group	0.002	[-0.005, 0.01]	1.00	0.004	.529
Block order (baseline = first)	-0.006	[-0.01, -0.003]	0.98	0.002	< .001***
Number of syllables in target segment	0.06	[0.05, 0.07]	1.15	0.01	< .001***
Log frequency of last two words	-0.004	[-0.02, 0.01]	0.99	0.01	.635
Log frequency of last three words	0.01	[-0.01, 0.04]	1.02	0.01	.324

Note. R^2 marginal = .45. R^2 conditional = .81. *SE* = standard error. NS = native English speakers. Log frequency of last word in the target phrase excluded to reduce collinearity among predictors.

3.5 Discussion

In this experiment, I set out to test whether adult native English speakers and ESL learners demonstrated frequency effects in compositional multi-word phrase production. As discussed in the literature review, research on frequency effects on L1 and L2 speech production durations was much more limited in comparison to research on comprehension. In particular, there had been only one relevant study with ESL learners (Ellis et al., 2008), but in that study frequencies of the subparts of the target phrases were not strictly controlled for. Moreover, no studies had investigated frequency effects in both comprehension and production in the same L1 and L2 speakers. As in previous relevant research (Arnon & Priva, 2013; Bannard & Matthew, 2008), the outcome variable in the present experiment is the first three words (e.g., *don't have to*) in the target four-word phrases (e.g., *don't have to worry*). In the analysis, frequencies of the subparts that differed in the two phrases in each pair were controlled for. The purpose was to ensure that any difference between the production durations of the target segments in the

high and low frequency phrases in each bin can be attributable to the difference in whole phrase frequency, not to the difference in subpart frequencies. The results from both cut-off bins suggested that participant group, stimuli block order, and the number of syllables of the target segments significantly predicted production durations of the target segments. With regard to the main effect of group, native English speakers' shorter production durations compared to those from ESL learners were not surprising, given the former group's higher English proficiency. The finding was in line with previous empirical evidence that speech fluency as measured objectively with syllable durations (as opposed to perceived fluency measured by human rating), increased as a function of language proficiency (De Jong, Groenhout, Schoonen, & Hulstijn, 2013; Kahng, 2014). In light of usage-based approaches (e.g., Bybee, 2010), the faster speech production in the native English speakers in the current study suggested that the connection between words in the target phrases in the native speakers' linguistic representation was stronger, leading to faster productive processing.

In the analysis for the current experiment, block order and the number of syllables of the target segments were control variables. It was possible that these two variables would influence production durations of the target segments, and based on the results, this possibility was confirmed. The block order effect was similarly reported in previous comprehension studies with a within-subject design with two counterbalanced experiment blocks (e.g., Arnon & Snider, 2010; Hernández et al., 2016) and in the first experiment in the current study. The main effects of block order in this second experiment therefore suggested that block order also affects phrase production durations. In previous studies on frequency effects on compositional phrase production (Bannard &

Matthew, 2008; Ellis et al., 2008; Tremblay & Tucker, 2011), target phrases were randomly presented and were not be divided into blocks; thus, the effect of block order did not surface as a factor impacting on production durations. In the current experiment, the two phrases from each pair were randomly put into two different blocks to reduce the repetition effect resulting from participants' production of the same target segment (e.g., *don't have to*) in the two phrases from a target pair (e.g., *don't have to worry* and *don't have to wait*) in the same experiment block. The finding suggested that if future speech production studies use a within-subject design with counterbalanced blocks to investigate frequency effects on phrase production durations, the effect of block order cannot be ignored. Otherwise, researchers may not be able to draw a conclusion that participants' production durations were driven by phrase frequency alone. With regard to the significant effects of the number of syllables, the results were in the expected direction; participants should need more time to produce a segment with more syllables. This was also congruent with findings from previous relevant research (Tremblay & Tucker, 2011).

In the current experiment, in both cut-off bins and in both participant groups, overall production durations of high frequency phrases were shorter than production durations for low frequency phrases (Table 10). However, based on the regression results, I did not find evidence that phrase frequency was a significant predictor of production durations for the target segments when the other differences between the two phrases in each pair were controlled for. Given previous relevant research, two possible reasons for the absence of phrase frequency effects could be the design and task instruction in the current experiment and the nature of the speech elicited. These are

discussed in the following sections.

The design of the current experiment differed from the design in previous research on frequency effects on production durations of compositional phrases beyond two words in a few ways. Previous research in this area can be broadly divided into two groups: (1) studies that used an elicited production task similar to the task in the current experiment (Arnon & Priva, 2013; Bannard & Matthew, 2008; Ellis et al., 2008; Tremblay & Tucker, 2011), and (2) studies that investigated frequency effects in spontaneous speech production (Arnon & Priva, 2013, 2014). Among the studies in the first group, three were conducted only with adult native speakers (Arnon & Priva, 2013; Bannard & Matthew, 2008; Tremblay & Tucker, 2011), while the other was conducted with both native English speakers and ESL learners (Ellis et al., 2008). The findings from the studies in this first group have been mixed. Ellis et al. (2008) reported frequency effects on production durations only from adult ESL learners, but not from adult native English speakers, and Tremblay and Tucker (2011) similarly did not find frequency effects from adult native English speakers. In these two studies, as in the current experiment, participants were instructed to say the target phrases as fast as they could in an elicited production task. However, as pointed out, in Ellis et al.'s (2008) study, substring frequencies were not strictly controlled for. Also, some of the target phrases were complete syntactic constituents (e.g., *a great deal of*), but some were not (e.g., *and at the, and so on but, the way in which the*). On the other hand, in the current experiment, substring frequencies were controlled for and the two phrases in each pair had the same constituency type (e.g., verb phrase, noun phrase). These methodological differences may have contributed to the dissimilarity between the results from Ellis et

al.'s (2008) experiment and the current experiment.

With regard to Tremblay and Tucker's (2011) study, their target phrases were four-word phrases derived from the BNC. In light of usage-based views of language acquisition, the predictor of interest was the frequency of the whole phrases. As in the current experiment, the researchers entered frequencies of the subparts of the target phrases as controll variables. However, the principle on which some other control variables were entered into Tremblay and Tucker's (2011) regression analysis was not always clearly explained. For example, the researchers included (1) the interaction between frequency of the first word and the frequency of the third word in the target four-word phrases, and (2) the interaction between whole phrase frequency and the frequency of the first two words in the target phrases. These interactions were significant in their regression results, but whether and how these interactions were meaningful were not clarified. While the current experiment and Tremblay and Tucker's (2011) study did not find evidence for frequency effects on production durations from native English speakers, given such methodological issues, it might still not be safe to conclude that the results from the current experiment and Tremblay and Tucker were completely compatible.

The current experiment was perhaps relatively more methodologically comparable with the elicited production experiments by Arnon and Priva (2013) and Bannard and Matthew (2008). These two experiments were conducted with only native English speakers, the former with adults and the latter with children aged 2-3 years old. As in the current experiment, the target stimuli in these experiments were pairs of four-word compositional phrases that differed only in the last word, and the target segments were also the first three words in the phrases. Moreover, an elicited production task was

also used, and substring frequencies were controlled for in the analyses. Unlike the current experiment, however, both of these experiments reported frequency effects on compositional multi-word phrase production durations. The incongruent results may have resulted from some remaining methodological differences. First, as in the current experiment, Arnon and Priva (2013) used a subset of the phrases from Arnon and Snider (2010) as stimuli. However, the researchers used a between-subject design to address a possible repetition effect resulting from a participant's reproduction of the identical first three words from a target pair (e.g., *don't have to worry* and *don't have to wait*). The two phrases in each pair were thus assigned to two different lists, and each of the 34 participants read only one of the lists. Therefore, one participant's production of *don't have to* in *don't have to worry* was compared against another participant's production of this same segment in *don't have to wait*. To account for the production speed variation among participants, Arnon and Priva (2013) entered the average production durations of each participant (across all target stimuli) in their regression model as a control variable and additionally entered the average production duration of each target segment (e.g., *don't have to*) (across all participants) as another control variable. In the current experiment, however, I attempted to minimize the possible repetition effect by assigning the two phrases in each pair to two different blocks separated by a break, and I controlled for individual variability by using a within-subject counterbalanced design. I also statistically controlled for the effects of block order in the regression analysis to eliminate the effects of task familiarity on production durations. I used this design to maintain consistency with the design in the phrasal acceptability judgment task in the first experiment, and I also believe that this design may be better at controlling for individual

variability because every participant did produce both phrases from each pair. However, it might have been these methodological dissimilarities that made Arnon and Priva's (2013) results and my results differ. Given the relatively small amount of research on frequency effects on compositional multi-word sequence production durations, this speculation should be investigated in future research.

In regard to the experiment by Bannard and Matthew (2008), which reported frequency effects in English speaking children, one important difference between that experiment and the current experiment was the direction in the task. Bannard and Matthew (2008) asked the children in their study to "say the same thing" (p. 44) after the children heard each target phrase from an audio clip. On the other hand, in the current experiment, participants were instructed to say the phrase as fast as they could while still being accurate after reading each phrase on a computer screen. Possibly, the instruction in the current experiment prompted the participants to be more focused on producing the phrases and therefore the difference between the production durations of high and low frequency phrases were less pronounced than that in Bannard and Matthew's (2008) study, especially because the unit of measurements of the production durations was as fine-grained as milliseconds. Moreover, in Bannard and Matthew's (2008) study, if children did not respond within a reasonable amount of time, the experimenter prompted them to respond once (e.g., by saying *Can you say that?*). By contrast, all participants in the current experiment had only one chance to respond, and they had to respond within the 2,500-millisecond interval immediately after reading each target phrase on a computer screen. In addition, possibly the children in Bannard and Matthew's (2008) study were not as attentive as the adults in the current experiment. Bannard and Matthew

(2008) asked each child to pronounce 32 phrases, one at a time, and retained only error-free productions in the analysis. Based on the responses from 17 two-year-old children and 21 three-year-old children, the researchers excluded all the data from the two-year old children because 68% of the responses contained errors and thus there were insufficient data for an analysis for this group. Bannard and Matthew's (2008) analysis was thus based only on production durations from the three-year-old children, but 34% of the data from this group also had to be excluded due to production errors. On the other hand, the mean production accuracy in the native speaker group and the ESL learner group in the current experiment was as high as 99% (*SD* = 0.01) and 97% (*SD* = 0.03) respectively. Possibly, the different amount of attention, which may have been reflected by the different accuracy rates, and the dissimilar instructions may have contributed to the incongruent results between the current experiment and Bannard and Matthew's (2008) experiment.

Interestingly, the only two existing studies on frequency effects on compositional multi-word sequence production durations in spontaneous speech of native English speakers similarly reported the effects—that is, the study by Arnon and Priva (2013) and a later study by Arnon and Priva (2014). In the former study, the target phrases were three-word sequences (e.g., *everybody was trying*, *saw the boy*) derived from the Switchboard corpus of spontaneous telephone conversations. In the latter study, the target phrases were three-word sequences in which the middle word is a noun, derived from the Buckeye Corpus of Conversational Speech (Pitt et al., 2007), which contained spontaneous interview speech. In both studies, substring frequencies, derived from the Fisher and Switchboard corpora, were entered as control variables (substring frequencies

were derived from two corpora combined for more reliability). The overall results seemed to support the psychological reality of frequency effects in adult native speakers' speech production durations. Arguably, the nature of speech investigated in these two experiments differed from the nature of the speech analyzed in the elicited production tasks (Arnon & Priva, 2013; Bannard & Matthew, 2008; Ellis et al., 2008; Tremblay & Tucker, 2011), including the task in the current experiment. In spontaneous production, it was conceivable that speakers have to productively process both form and meaning in a natural conversational-like fashion, while in an elicited production task, participants may focus only on repeating the sentences they saw on a computer screen or heard from an audio clip. As discussed, the results from elicited production experiments have been mixed, and the reason for the mixed results could have been cross-study methodological differences. However, given the results from the spontaneous speech production research (Arnon & Priva, 2013, 2014), another possibility may be that frequency effects on phrase production durations may be more likely to be observed in more natural, spontaneous speech. Interestingly, frequency effects on production durations were also reported in previous research focusing on spontaneous production of single words (e.g., Bell et al., 2009; Bybee & Scheibman, 2000; Jurafsky et al., 2000). The use of the elicited production task in the current experiment was motivated by the support for frequency effects in previous elicited phrase production experiments (Arnon & Priva, 2013; Bannard & Matthew, 2008). Moreover, because in the current study, I investigated frequency effects on both comprehension and production in the same participants, the use of the elicited production task allowed me to investigate processing of the same target phrases in both comprehension and production and therefore to maximize the

comparability of the two tasks. Therefore, whether or how the different nature of speech analyzed (elicited vs spontaneous) contributes to a presence or an absence of frequency effects on compositional multi-word sequence production durations is worthy of investigation in future research.

CHAPTER 4: GENERAL DISCUSSION

In the current study, I investigated whether frequency effects can be observed in the processing of multi-word compositional processing in adult native English speakers and ESL learners. Such effects are predicted in usage-based approaches to L1 and L2 acquisition (e.g., Ambridge et al., 2015; Arnon, 2015; Bybee, 2010; Ellis, 2011; Ellis et al., 2013; Ellis & Larsen–Freeman, 2009; Gries & Ellis, 2015; Tomasello, 2003, 2009). The research questions are (1) whether speakers in these two groups demonstrate sensitivity to frequency of compositional four-word English phrases in recognition when the frequencies of the smaller parts are controlled for, and (2) whether the speakers also demonstrate such sensitivity during language production when the frequencies of the smaller parts are controlled for. Two separate respective experiments were conducted with the same participants to answer these two questions. The current study is the first that investigated frequency effects on such word sequences in both comprehension and production in the same L1 and L2 speakers.

In the first experiment, receptive processing was operationalized as reaction times in a phrasal acceptability task, in which participants judged whether the target phrases were possible English word sequences. The participants in both groups demonstrated frequency effects. That is, they reacted faster to higher frequency phrases than to lower frequency phrases. The effects in the native English speakers were compatible with results from the two previous relevant studies with substring frequency control—namely, the studies by Arnon and Snider (2010) and Hernández et al. (2016). The collective results and the existing findings that in receptive processing native English speakers were sensitive to frequency of inflectional morphemes (see Ambridge et al., 2015 for a

review), single words (e.g., Diependaele et al., 2012; Whitford & Titone, 2012), idioms (Nippold & Rudzinski, 1993), and two-word compositional phrases (e.g., Gyllstad & Wolter, 2016; Wolter & Gyllstad, 2013) provide empirical evidence supporting usage-based researchers' claim that frequency is likely to influence the representation and processing of L1 linguistic units at all levels. The present study was among the first few studies which helped fill an empirical gap because other existing L1 receptive studies either focused on two-word compositional phrases (e.g., Sonbul, 2015; Sosa & Macfarlane, 2002; Wolter & Gyllstad, 2013) or did not controlled for frequencies of subparts of the target phrases (e.g., Ellis et al., 2008; Siyanova–Chanturia, Conklin, & van Heuven, 2011; Tremblay et al., 2011). That is, with substring frequency control, frequency effects on four-word phrase comprehension provided stronger support for usage-based approaches. As Arnon (2015) pointed out, "Frequency effects are not interesting in and of themselves. They are interesting because they reveal something about the [language] learning mechanisms …" (p.274). Multiword frequency effects suggest that the domain-general human cognitive processes, such as chunking and rich memory, and the accumulation of statistical information in previously-encountered input do affect L1 representation and processing (e.g., Abbot–Smith & Tomasello, 2006; Bybee, 2010; Ellis, 1996, 2002, 2011; Gries & Ellis, 2015). Moreover, frequency effects on L1 compositional multi-word phrase processing seemed incompatible with an L1 language acquisition or processing model in which the lexicon and grammar are completely divided and in which frequency effects should be observed only with items in the mental lexicon (i.e., morphemes, words, idioms), but not with compositional phrases,

which were generated based on abstract grammar rules (e.g., Prasada et al., 1990; Prasada & Pinker, 1993; Ullman, 1999).

With regard to ESL learners, the results from the first experiment in the current study, together with the recent findings from Hernández et al. (2016), extend the preceding empirical support for frequency effects on ESL receptive processing of adjective-noun or verb-noun collocations (e.g., Gyllstad & Wolter, 2016; Sonbul, 2015; Wolter & Gyllstad, 2013) to longer compositional sequences. These therefore appeared to corroborate usage-based researchers' proposal that the general cognitive processes that operate in L1 acquisition may also operate in L2 acquisition (e.g., Ellis, 2008a, 2008b, 2012, 2013; Ellis & Cadierno, 2009; Ellis & Larsen–Freeman, 2009; Ellis & Wulff, 2015; Goldberg & Casenhiser, 2008; Robinson & Ellis, 2008). The cumulative results, based on ESL learners from different L1 backgrounds, also seem to counter Wray's (2002) hypothesis that adult L2 learners cannot retain in their memory information about L2 word co-occurrences. Wray's proposal (2002) was that L2 learners are likely to break previously-encountered frequent word sequences into individual words due to a lack of necessity to use these sequences, their L2 education, which typically focuses on forms and individual words, and their mature cognitive development and L1 literacy. If Wray's (2002) claim is true, adult ESL learners' compositional word sequence processing should not have been frequency sensitive. In addition, previous studies have reported frequency effects on ESL single word recognition (e.g., Diependaele et al., 2012; Whitford & Titone, 2012). Consequently, there seems to be evidence supporting usage-based researchers' proposal that frequency may affect the acquisition and processing of L2 linguistic units at all levels, not just single words; that is, the representation and

processing of L2 words and compositional phrases are under a similar mechanism (e.g., Bybee, 2008; Ellis, 2011; Ellis & Cadierno, 2009; Ellis & Larsen–Freeman, 2009; Ellis & Robinson, 2008; Ellis & Wulff, 2015; Eskildsen, 2012; Goldberg & Casenhiser, 2008; Römer et al., 2014).

As discussed in Chapter 2, the results from the first experiment in the current study also suggested that participants in both groups processed high frequency phrases in both stimuli cut-off bins faster. This suggested that participants did not process only compositional phrases with very high frequency (i.e., high frequency phrases in the high cut-off bin) faster; similar frequency effects were observed in the low cut-off bin. The results support the hypothesis that differences in reaction times to the target phrases should have resulted from relative quantitative differences (i.e., different frequencies of previous phrase encounters) regardless of the frequency range (Bybee, 2006, 2010; Bybee & Hopper, 2001). That is, as in the study by Arnon and Snider (2010), the results do not support the hypothesis that there is a high frequency threshold (e.g., 12 occurrences per million words), and only a phrase with a frequency above this threshold is stored as a whole and is processed faster than a less frequent phrase, which is not stored as a whole but is computed or analyzed based on grammar (e.g., Wray, 2002)

Regarding the second research question, in line with previous research (Bannard & Matthew, 2008; Arnon & Priva, 2013), productive processing was operationalized as production durations of the first three words (e.g., *don't have to*) in the four-word target phrases (e.g., *don't have to worry*) in an elicited production task. Another reason for the use of this task was because it allowed me to investigate frequency effects on receptive and productive processing of the same target phrases, thus allowing me to maximize the

comparability between the two experiments in the current study. All participants in the current study completed the receptive task before the production task; the purpose was to avoid the possibility that participants' exposure to the target phrases in the production task affected their judgments in the receptive task. Moreover, one source of the study motivation was that no previous research on compositional multi-word sequence processing simultaneously looked at frequency effects on both comprehension and production in the same participants with substring frequency control, and such research may reveal whether task characteristics (comprehension vs production) could impact frequency effects. Moreover, relevant L2 production research has been very limited (Siyanova–Chanturia & Martinez, 2014), and the only existing study, conducted by Ellis et al. (2008), did not control for substring frequency.

As noted, previous studies with native English speakers reported an inverse relationship between frequency and production durations of single words. For example, Bybee and Scheibman (2000) reported that the production duration of *don't* in spontaneous interview conversations was shorter when it was embedded in frequent phrases, such as *I don't know* (contracted forms such as *don't* are also generally considered a single word in speech production studies). Similarly, Bell et al.'s (2009) analysis of word durations in the Switchboard corpus of telephone conservations suggested that word frequency negatively and significantly predicted production durations of both content and function words. Bell et al. (2003) and Jurafsky et al. (2000) additionally found that words embedded inside more frequent two-word sequences in Switchboard were produced faster than those in less frequent sequences. Despite these results, additional evidence for usage-based approaches will be from production of

phrases consisting of more than two words. Based on usage-based researchers' claim about L1 and L2 acquisition and based on the results from the single word production research, it seems reasonable to hypothesize that, based on production durations, more frequent compositional phrases will be productively processed faster by both native English speakers and ESL learners (Arnon & Priva, 2013, 2014; Ellis et al., 2008).

However, while the participants in both groups in the current study had shorter average production durations for high frequency phrases than for low frequency phrases, these participants, who demonstrated frequency effects in the comprehension task, did exhibit frequency effects in the elicited production task when substring frequencies of the target phrases and other relevant factors that may have affected production durations (i.e., block order and the number of syllables of the target segments) were controlled for. Results from the existing research based on elicited production of multiword sequences, including the second experiment in the current study, have been mixed. Some studies found frequency effects in native English speakers (Arnon & Priva, 2013; Bannard & Matthew, 2008), while some did not (Ellis et al., 2008; Tremblay & Tucker, 2011). In the case of ESL learners, besides the current study, Ellis et al.'s (2008) study is the only existing study and reported the frequency effects. As discussed, several cross-study methodological differences could have contributed to these incongruent findings, and such differences should be investigated in future studies, particularly because there has been relatively limited research on frequency effects on compositional multi-word sequence production durations, especially in ESL learners.

One additional observation that I pointed out earlier is that, unlike previous research based on an elicited production task, the existing research on compositional

multi-word sequence production based on spontaneous speech from native English speakers consistently reported frequency effects (Arnon & Priva, 2013, 2014). These more consistent findings, together with the findings from L1 spontaneous word production duration research (e.g., Bell et al., 2003, Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Bybee & Scheibman, 2000, Jurafsky et al., 2000), suggest that the type of speech investigated (elicited vs spontaneous) may be another factor contributing to whether frequency effects can be observed. Possibly, frequency effects on compositional multi-word sequences will be observed more consistently in future research if spontaneous speech from native English speakers and ESL learners is the subject of investigation. This will be particularly interesting in the case of ESL learners because to date frequency effects on compositional multi-word sequence production in spontaneous ESL speech with substring frequency control have never been investigated.

Finally, although my decision to use an elicited production task was informed by previous studies (Bannard & Matthew, 2008; Arnon & Priva, 2013; Ellis et al., 2008; Tremblay & Tucker, 2011), one limitation in the current study seems to be that the phrasal decision task in the first experiment and the elicited production task in the second experiment are not completely comparable. In the phrasal decision task, participants had to process both the form and the meaning of the stimuli to judge whether the word sequences were possible English phrases. There were also ungrammatical sequences, which were distractors, and therefore the participants had to be careful when making their decisions. By contrast, in the elicited production task, the participants merely repeated the target phrases accurately. If frequency effects in phrase production are related to "activation of multi-word lemmas" (Arnon & Priva, 2013), it might be possible that by

the time each target phrase disappeared from the computer screen in Experiment II, the multi-word lemmas had already been activated, so the difference in production durations between high and low frequency phrases was reduced. Based on the existing relevant research and the current study, more production studies, particularly those based on spontaneous speech, could be conducted to investigate whether usage-based researchers' claim can be corroborated with evidence based on phrase production durations, particularly in L2 learners.

CHAPTER 5: CONCLUSION

5.1 Summary of findings

Motivated by usage-based approaches to L1 and L2 acquisition, the current study is the among the studies which investigated frequency effects on the processing of compositional phrases beyond two-words in adult L1 speakers (Arnon & Snider, 2010; Arnon & Priva, 2013, 2014; Bannard & Matthew, 2008; Tremblay & Tucker, 2011) and L2 speakers (Hernández et al., 2016) with an attempt to control for frequencies of the subparts of the target sequences. The current study also constituted the first effort to investigate such effects in both language comprehension and production in the same L1 and L2 speakers.

The results from the phrasal acceptability task suggested phrase frequency effects in adults native English speakers and ESL learners who were sufficiently proficient in English to study in a US university environment and who had lived in the US for approximately 2-3 years. These were in line with findings from the existing relevant L1 and L2 receptive research (Arnon & Snider, 2010; Hernández et al., 2016) and supported usage-based views of language acquisition. Moreover, because frequency effects were observed in both the high and low cut-off bins, the results support the proposal that the differences in reaction times to the target phrases should have resulted from relative phrase frequency differences (Bybee, 2010; Bybee & Hopper, 2001). Such a result pattern was observed from both native English speakers and the ESL learners. Overall, these were in line with previous findings from Arnon and Snider (2010) and Hernández et al. (2016). Further, results from the first experiment suggest that frequency data from a large native English corpus—or specifically the 20-million-word corpus based on the

Fisher and Switchboard corpora combined—seem to be representative of the general regularities of English input that both the native English speakers and ESL learners in the current study had been exposed to.

On the other hand, in the elicited phrase production task, I did not find evidence for frequency effects on production durations of compositional multi-word sequences. The results from the existing relevant research based on phrase production durations, including the second experiment in the current experiment, have been mixed. The inconsistent findings could have resulted from several cross-study methodological differences, such as the type of instruction (Bannard & Matthew, 2008), the lack of control of substring frequency and syntactic constituency (Ellis et al., 2008), the principle on which some control variables were entered into the analysis (Tremblay & Tucker, 2011), and whether the study has a within or a between subject design (Arnon & Priva, 2013). As a result, this possibility could be investigated in future research. In addition, while informed by previous relevant studies, the two experiments in the current study were arguably not completely comparable. In the first experiment, participants had to carefully process both the form and the meaning of the target phrases, while in the elicited production experiment, they only focused on repeating the target phrases correctly. This task difference may be another reason why the participants in both groups appeared to demonstrate frequency effects in the first but not in the second experiment. Finally, given the more consistent empirical support for frequency effects on production durations of L1 words (e.g., Bell et al., 2003, 2009; Bybee & Scheibman, 2000) and L1 compositional multi-word sequences (Arnon & Priva, 2013, 2014) in spontaneous speech, possibly frequency effects on L1 and L2 multi-word phrases will be observed

more consistently if future research focuses on spontaneous phrase production durations, although creating or finding a sufficiently large L2 corpus for such research may constitute a practical challenge.

5.2 Implications

The results from the first experiment in the current study suggested native English speakers' sensitivity to frequency during comprehension of compositional multi-word phrases. Given such findings and the other documented frequency effects on receptive processing of various types of L1 linguistic units—including words, inflectional morphemes, and compositional phrases—in both children and adults (e.g., Arnon, 2015; Arnon & Snider, 2010; Diependaele et al., 2012; Diessel, 2007; Ellis, 2002; Hernández et al., 2016), the next question is how to develop a plausible relevant L1 representation and processing psycholinguistic model with descriptive and predictive adequacy (Ibbotson, 2013). In this regard, usage-based researchers seem to have relied on simulated computer models constructed based on some general principles compatible with usage-based views—that is, language acquisition and processing result from an accumulation of statistics information in previously-encountered input; the lexicon and grammar are not rigidly divided; and the processes underlying the operation of the models are analogous to domain-general human cognitive processes. According to Ellis (2011), these models include connectionist and statistical learning models (Bod, Hay, & Jannedy, 2003; Christiansen & Chater, 2001; Elman et al., 1996; MacWhinney, 1992, 1997; Rumelhart & McClelland, 1986). Arnon and Snider (2010) emphasized that the existing L1 speech perception models not only have to be expanded to accommodate multi-word frequency, but also have to account for the complex relationships between linguistic units at varying

degrees of complexity and abstraction. For example, as noted, in usage-based views, multiple encounters of lexically specific sequences (e.g., *David kissed Liz*) lead to subsequent acquisition of more abstract related sequences (e.g., *KISSER kissed KISSEE, David kissed KISSEE, Subject-Verb-Object),* and speakers can simultaneously store a fully lexically filled sequence and related more abstract sequences (e.g., Ambridge & Lieven, 2011). Therefore, as pointed out by Arnon and Snider (2010) and Ibbotson (2013), an adequate model has to address several key issues, such as (1) how encounters of lexical sequences lead to an inference about the underlying abstract forms, (2) how an encounter of a specific lexical sequence is counted as an instance of multiple more abstract sequences, and (3) the relationship between the representation of multi-word phrases (e.g., *don't have to worry*), the subparts (e.g., *to worry*), and the more abstract linguistic units related to the subparts (e.g., an infinitive clause). Such a model will rely on processes such as chunking and categorization (e.g., Bybee, 2010) and will have to be able to expand and organize itself internally. Given such complexity, in a relatively recent review of the scope of usage-based views of language and acquisition, Ibbotson (2013) conceded that developing a plausible computer model with sufficient explanatory and predictive power is still a challenge for usage-based researchers.

In terms of language production, the elicited production task in the current study did not suggest frequency effects on L1 compositional multi-word sequence production durations. There is however increasing evidence in the L1 literature for frequency effects on single word and compositional multi-word sequence production in on spontaneous speech (e.g., Bybee & Scheibman, 2000; Bell et al., 2009; Arnon & Priva, 2013, 2014). As discussed, given the mixed findings from the existing research based on an elicited

production task, it may be possible that frequency effects on production durations of compositional multi-word sequences will be documented more consistently if future research focuses on spontaneous speech. Sensitivity to frequency in compositional phrase production will entail the need for a development of a psycholinguistic model that accommodates both word and multi-word phrase frequency in speech production. The model also has to account for several usage-based proposals, such as the assumption that a speaker's utterances (e.g., *David kissed Liz*) are not always formed from the most abstract stored representation possible (e.g. *Subject-Verb-Object)*, but, depending on the speakers' previous linguistic experience, may be formed from any combinations of abstract and more concrete relevant components (e.g., *KISSER kissed KISSEE, David kissed KISSEE)* (Ambridge & Lieven, 2011). Arnon and Priva (2013) suggested that such a model can be an expanded version of the connectionist models of L1 production (e.g., Chang, 2002; Chang, Dell, & Bock, 2006). However, as the two researchers observed, developing an adequate model, in which there are activation and competition among not only single words but also multi-word phrases, is a challenge yet to overcome.

Similarly, with regard to L2 learners, the increase in empirical support for frequency effects, though sometimes at various degrees, on receptive processing of ESL compositional phrases (Gyllstad & Wolter, 2016; Hernández et al., 2016; Sonbul, 2015; Wolter & Gyllstad, 2013), including the effects in the first experiment in the current study, necessitate a development of a relevant L2 psycholinguistic model. It is conceivable the model will be more complicated than an L1 model. For example, L2 studies have suggested various types of influence that an L1 may exert on L2 compositional phrase acquisition and processing (e.g., Römer et al., 2014; Wolter &

Gyllstad, 2013). Another relevant factor could be L2 learners' English proficiency. Hernández et al. (2016) pointed out that the connectionist computer models of L2 acquisition (MacWhinney, 2008) can be expanded to accommodate multiword frequency. In addition and as noted, the experiment by Ellis et al. (2008) and the second experiment in the current study are the only two existing experiments that attempted to shed light on frequency effects on compositional multi-word phrase production durations in L2 speakers, and the latter experiment is the first L2 experiment that attempted to control for frequencies of the subparts of the analyzed phrases. As in the case of L1 learners, more relevant studies with L2 spontaneous production should be conducted. In sum, evidence for L1 and L2 multi-word frequency effects will require a future development of plausible psycholinguistic models with descriptive and predictive power that attest the ontological status of the theoretical claims in usage-based approaches.

5.3 Limitations and future research

The current study has some limitations that future research could address. First, the two research questions concerned the processing of English compositional multi-word sequences, but the target phrases constituted a specific subset of such sequences. All the target phrases were four-word phrases; moreover, in each pair, the two phrases had identical first three words and thus differed only in the last word. This allowed for an easier control of substring frequency in the regression analyses. That is, if there had been more different subparts, more substring frequencies must have been controlled for. For example, if the two phrases in each pair shared only the first two words (e.g., *don't have to worry* vs *don't have any sisters*), there would have been six different subparts: the third

word (*to* vs *any*), the last word (*worry* vs *sisters*), the second two words (*have to* vs *have any*), the last two words (*to worry* vs *any sisters*), the first three words (*don't have to* vs *don't have any*), and the last three words (*have to worry* vs *have any sisters*). Future studies using a wider range of L1 and L2 stimuli will shed light on whether the results in the current study can be generalized to other type of compositional phrases consisting of more than two words.

Second, unlike previous researchers investigating ESL compositional phrases beyond two words (Ellis et al., 2008; Hernández et al., 2016; Siyanova–Chanturia, Conklin, & van Heuven, 2001; Valsecchi et al., 2013), I controlled for participants' L1 and substring frequencies. I also ensured that all ESL learners knew all the component words in the target phrases. However, one issue I did not address was whether each target phrase has a direct word-for-word L1 translation and whether such a translation and a lack thereof affected the ESL learners' phrase processing. Previous ESL adjective-noun and verb-noun collocation studies have distinguished between two collocation types: congruent and incongruent (Wolter & Gyllstad, 2013; Yamashita & Jiang, 2010). A congruent L2 collocation has a direct word for word L1 translation equivalent. For example, according to Yamashita and Jiang (2010), *hot tea* is a congruent L2 collocation for Japanese learners of English because the corresponding word for word translation was a possible Japanese phrase. By contrast, *strong tea* is an incongruent L2 collocation; the semantic concept was expressed with Japanese phrases that translate as *dark tea*, *dense tea*, or *thick tea* in English. Using a phrasal acceptability task, Yamashita and Jiang (2010) investigated receptive processing of adjective-noun collocations (e.g., *heavy stone*) and verb-noun collocations (e.g., *make lunch*) in two groups of Japanese English

learners: (1) Japanese students, researchers, or instructors at a US university, and (2) university students in Japan. The results revealed that congruency did not significantly predict reaction time from participants in the first group. However, participants in the second group, with no experience living in an English speaking country and with less English exposure, processed congruent L2 collocations significantly faster than incongruent collocations. This may therefore suggest that congruency affects reaction time only in the initial state of L2 learning. Yamashita and Jiang (2010) contended that the results were compatible with Jiang's (2000) prediction that L2 learners can more easily understand the meaning of congruent L2 collocations because of readily available L1 counterparts (i.e., L1 boost). By contrast, in the case of incongruent L2 collocations, L2 learners first need to comprehend the meaning of the component words. This may require contextual cues and thus establishing the L2 form-meaning connection is more challenging. In another English adjective-noun collocation study, Wolter and Gyllstad (2013) reported that Swedish learners of English enrolled in English courses at a university in Sweden processed congruent L2 sequences significantly faster than incongruent sequences, but regardless of the congruency status, frequency significantly predicted participants' reaction time. Wolter and Gyllstad (2013) argued that, based on Eurocentres Vocabulary Size Test (Meara & Jones, 1990), their participants were relatively advanced English learners, but acknowledged the difficulty in comparing their participants' English proficiency against the proficiency of ESL learners in the L2 literature because the test was not a common ESL proficiency measure. Therefore, in light of this observation, it was not clear whether the participants in Wolter and Gyllstad's (2013) study were similar to the more or the less advanced English learner

group in Yamashita and Jiang's (2010) study. The participants in the current study, ESL learners enrolled at a US university, had similar characteristics to the more advanced learner group in Yamashita and Jiang's (2010) study. Therefore, as in the case of those advanced learners, it may be possible that for the ESL learners in the current study, congruency did not significantly affected reaction times. Another possibility may be that congruency plays a smaller role in the processing of compositional sequences in the current study because the component words in the target phrases generally convey a literal meaning (e.g., *don't have to worry*), unlike many phrases in Yamashita and Jiang's (2010) and Wolter and Gyllstad's (2013) studies in which a component word conveyed a figurative meaning (e.g., *kill time, strong tea, break a promise*). A recent study by Gyllstad and Wolter (2016) reported that advanced English learners receptively processed phrases in which all component words convey a literal meaning (e.g., *kick a ball*) significantly faster than phrases in which a component word conveys a figurative meaning (e.g., *draw a conclusion*), but regardless of the semantic transparency of the component words, phrase frequency was a significant predictor of participants' reaction time. However, all of Gyllstad and Wolter's (2016) stimuli were congruent L2 sequences. As a result, the interaction between congruency, semantic transparency of component words, and frequency effects on compositional multi-word sequence comprehension and production could be further explored in future research. It is also noteworthy that in previous ESL research investigating the role of L1-L2 congruency, it was not always clear if participants knew all the component words in the target phrases. Yamashita and Jiang (2010), for example, conceded this limitation. Thus, future studies

in this area should control for word familiarity because it may affect participants' reaction time and thus constitute an experimental confound.

In addition, in the case of English learners, one possible issue that could be explored further is whether frequency effects on compositional phrase processing is influenced by English proficiency. In previous comprehension studies, some researchers did not find that ESL proficiency mediates frequency effects. For example, in Sonbul's (2015) reading task with concurrent eye-movement registration, the researcher found the effects of English adjective-noun collocation frequency on adult ESL learners' first pass reading time, but there was no significant interaction between collocation frequency and ESL proficiency, as measured by the Vocabulary Levels Test (Meara & Jones, 1988), in their regression analysis. Similarly, using reaction time in a phrasal acceptability task as the outcome measure, Hernández et al. (2016) did not find that ESL proficiency, as measured by Lexical Test for Advanced Learners of English (Lemhöfer & Broersma, 2012), modulated phrase frequency effects. Likewise, in Siyanova–Chanturia, Conklin, and van Heuven's (2011) eye-tracking study, ESL learners demonstrated sensitivity to phrase frequency, but there was no interaction between phrase frequency and the learners' self-reported English proficiency, which was based on a 5–point Likert scale (1 = poor, 5= excellent). Hernández et al. (2016) and Sonbul (2015) speculated that in their studies there may not be sufficient variation in the ESL proficiency among the participants for the interaction between ESL proficiency and phrase frequency to emerge. That is, the recruited participants were relatively advanced English learners—students enrolled at a university in an English speaking country (Hernández et al., 2016; Sonbul, 2015) or students majoring in a language-related field (Hernández et al., 2016). The

researchers therefore argued that such an interaction may be observed in future studies if there is a wider range of ESL proficiency among participants. In the current study, I did not manipulate English proficiency as another explanatory variable. While trying to minimize participants' individual variability in terms of their previous length of stay in an English environment, I recruited ESL learners who should not have difficulty understanding the target phrases, and my minimum TOEFL and IELTS scores were also slightly higher than the minimum scores in several previous studies on ESL word sequence memory retention or comprehension (e.g., Sonbul, 2015). However, the role of ESL proficiency on frequency on the processing of compositional multi-word sequences is an interesting issue that future research can investigate.

Further, the results from the current study and the recent study by Hernández et al. (2016), which were based on target phrases from Arnon and Snider (2010), suggested that, when reaction time in phrasal acceptability judgment tasks was the outcome measure, both native English speakers and ESL learners were sensitive to phrase frequency and that the former group processed English phrases faster. However, both studies did not find an interaction between frequency and participant group. I previously discussed that these differed from findings in single word recognition research, in which frequency effects were observed from both participant groups but were greater in L2 speakers, as indicated by the frequency by group interaction. Therefore, whether such an interaction does exist could be explored in future research. In addition, in the present study, frequency effects were not observed in the elicited production task, and there was no interaction between frequency and participant group as well. Therefore, as in the case of phrase recognition, future phrase production research could also investigate such an

interaction. A similar reason was that previous single word production research reported frequency effects in both L1 and L2 speakers, but greater effects in L2 speakers, as also indicated by the frequency and group interaction. This interaction has prompted researchers to propose *the weaker links hypothesis* (e.g., Gollan, Montoya, & Werner, 2002; Gollan et al., 2008), which posits that, due to less L2 exposure, there is a weaker link between semantics and phonology in L2 speakers' lexical system than in native speakers' system. Consequently, L2 word production requires more time, and disproportionally more time when L2 words have low frequency. The difference between productively processing high and low frequency words is therefore more pronounced in L2 speakers than in L1 learners. In sum, it remains to be seen whether an interaction between frequency and participant group will be observed in future studies on compositional multi-word sequence comprehension and production. The results could shed light on whether and how word and phrase processing are similar or different.

Finally, although frequency is a key index of linguistic experience (e.g., Ellis, 2002; Gries & Ellis, 2015) and is the only focus of the current study, in usage-based approaches to language acquisition, there are several other statistical information in previously-encountered input that may play a role in language acquisition and that speakers of a language may be sensitive to, such as mutual information, t-scores (Evert, 2008; Gries, 2010), delta P (Ellis, 2007; Ellis & Ferreira–Junior, 2009a; Gries, 2015), and word predictability (e.g., Bell et al., 2009). Such statistical information may also interact with frequency. For example, in the current study, each phrase was presented out of context and the only words that differed in each pair (e.g., *worry* vs *wait*) were content words. In light of previous work by Bell et al. (2009), Arnon and Snider (2010)

suggested that future research could also investigate if frequency effects are mediated by the type of words that differ in each pair when target phrases are contextualized. That is, Bell et al. (2009) found that native English speakers productively processed both a content word and a function word faster when the word is more predictable given a following word (i.e., when the conditional probability of the word in question is higher given the presence of the following word). By contrast, only the processing of very frequent function words was faster when the word is more predictable given a preceding word. Such a possible complex interaction between frequency effects, type of words, and word predictability given the context is an example of issues that can be investigated further.

APPENDICES

APPENDIX A: The 28 target pairs

The two tables below show the target pairs of phrases and their frequencies (F) per million words. In the high cut-off bin, based on the frequency I could derive from the Fisher and the Switchboard corpora, which together contain about 20 million words, the high frequency variants occur at least 12.00 times per million word, while the low frequency variants occurred less frequently. On the other hand, in the low cut-off bin, the cut-off frequency point between high and low frequency variants in each pair is 1.00 times per million word.

Table 13. Target pairs in the high cut-off bin

No.	Phrases	F	No.	Phrases	F
1.	a lot of places	12.80	9.	I have a lot	33.75
	a lot of days	0.70		I have a little	11.25
2.	a lot of work	19.25	10.	I have to say	21.00
	a lot of years	2.55		I have to see	1.40
3.	all over the place	27.05	11.	I want to go	12.80
	all over the city	0.85		I want to know	3.90
4.	don't have to worry	20.35	12.	It's kind of hard	17.10
	don't have to wait	2.00		It's kind of funny	9.05
5.	don't know how much	16.90	13.	on the other hand	36.70
	don't know how many	10.15		on the other end	4.80
6.	go to the doctor	19.70	14.	out of the house	12.00
	go to the beach	6.95		out of the game	0.80
7.	how do you feel	36.95	15.	we have to talk	12.60
	how do you do	6.60		we have to say	0.90
8.	I don't know why	47.85	16.	where do you live	53.15
	I don't know who	11.60		where do you work	4.35

Table 14. Target pairs in the low cut-off bin

No.	Phrases	F	No.	Phrases	F
1.	a lot of rain	6.00	7.	I want to say	5.60
	a lot of blood	0.25		I want to sit	0.35
2.	don’t have any money	2.80	8.	it was really funny	3.90
	don’t have any place	0.45		it was really big	0.20
3.	going to come back	1.85	9.	out of the car	2.60
	going to come down	0.55		out of the box	0.30
4.	have to be careful	7.10	10.	we have to wait	1.85
	have to be quiet	0.15		we have to leave	0.35
5.	I have a sister	6.95	11.	we have to talk	12.60
	I have a game	0.05		we have to sit	0.25
6.	I have to pay	2.80	12.	you like to read	2.10
	I have to play	0.15		you like to try	0.15

APPENDIX B: Frequency (F) of the target phrases and their sub-parts

In each phrase (e.g., *a lot of rain*), unigram1, unigram2, unigram3, and unigram4 refer to the first (e.g., *a*), second (e.g., *lot*), third (e.g., *of*), and fourth word (e.g., *rain*), respectively. Bigram1, bigram2, and bigram 3 are the first two words (e.g., *a lot*), the middle two words (e.g., *lot of*), and the final two words (e.g., *of rain*). Trigram1 and trigram 2 mean the first three words (e.g., a lot of) and the last three words (e.g., *lot of rain*) in each phrase. These were all the subparts in the target phrases.

Table 15. Frequencies of the target phrases in the low cut-off bin and their subparts

Pair	Type	Phrases	fphrase	fphrase/ million	Funigram1	Funigram2	Funigram3	Funigram4	Fbigram1	Fbigram2	Fbigram3	Ftrigram1	Ftrigram2
1	hi	a lot of rain	120	6	518,226	64,812	358,463	1,455	60,720	44,956	232	42,688	133
	lo	a lot of blood	5	0.25	518,226	64,812	358,463	1,097	60,720	44,956	37	42,688	6
2	hi	don't have any money	56	2.8	179,461	208,841	30,319	24,096	14,806	4,824	406	1,756	98
	lo	don't have any place	9	0.45	179,461	208,841	30,319	10,165	14,806	4,824	248	1,756	20
3	hi	going to come back	37	1.85	55,618	513,572	17,910	27,406	33,261	3,060	1,767	330	367
	lo	going to come down	11	0.55	55,618	513,572	17,910	22,442	33,261	3,060	466	330	88
4	hi	have to be careful	142	7.1	208,841	513,572	120,385	941	36,437	39,300	393	3,460	241
	lo	have to be quiet	3	0.15	208,841	513,572	120,385	673	36,437	39,300	74	3,460	21
5	hi	i have a sister	139	6.95	137,715	208,841	518,226	3,470	32,615	33,649	313	8,385	153
	lo	i have a game	1	0.05	137,715	208,841	518,226	5,958	32,615	33,649	975	8,385	14
6	hi	i have to pay	56	2.8	137,715	208,841	513,572	10,302	32,615	36,437	3,238	3,864	1,168
	lo	i have to play	3	0.15	137,715	208,841	513,572	10,141	32,615	36,437	2,307	3,864	52
7	hi	i want to say	112	5.6	137,715	30,722	513,572	39,886	4,239	18,081	7,537	2,571	347
	lo	i want to sit	7	0.35	137,715	30,722	513,572	4,924	4,239	18,081	1,050	2,571	91
8	hi	it was really funny	78	3.9	653,604	191,328	119,182	10,916	50,938	3,992	612	1,732	110
	lo	it was really big	4	0.2	653,604	191,328	119,182	21,112	50,938	3,992	571	1,732	17
9	hi	out of the car	52	2.6	74,497	358,463	666,006	5,980	13,508	49,334	1,216	2,890	95
	lo	out of the box	6	0.3	74,497	358,463	666,006	860	13,508	49,334	115	2,890	13
10	hi	we have to wait	37	1.85	186,179	208,841	513,572	3,408	15,528	36,437	711	2,288	345
	lo	we have to leave	7	0.35	186,179	208,841	513,572	3,890	15,528	36,437	1,023	2,288	138
11	hi	we have to talk	252	12.6	186,179	208,841	513,572	17,366	15,528	36,437	7,018	2,288	493
	lo	we have to sit	5	0.25	186,179	208,841	513,572	4,924	15,528	36,437	1,050	2,288	139
12	hi	you like to read	42	2.1	845,026	340,007	513,572	9,667	6,182	11,982	2,220	1,096	241
	lo	you like to try	3	0.15	845,026	340,007	513,572	12,476	6,182	11,982	2,168	1,096	113

Table 16. Frequencies of the target phrases in the high cut-off bin and their subparts

Pair	Type	Phrases	fphrase	fphrase/ million	Funigram1	Funigram2	Funigram3	Funigram4	Fbigram1	Fbigram2	Fbigram3	Ftrigram1	Ftrigram2
1	hi	a lot of places	256	12.8	518,226	64,812	358,463	5,323	60,720	44,956	463	42,688	271
	lo	a lot of days	14	0.7	518,226	64,812	358,463	7,179	60,720	44,956	591	42,688	14
2	hi	a lot of work	385	19.25	518,226	64,812	358,463	28,315	60,720	44,956	1,249	42,688	407
	lo	a lot of years	51	2.55	518,226	64,812	358,463	26,187	60,720	44,956	1,945	42,688	53
3	hi	all over the place	541	27.05	95,708	25,757	666,006	10,165	2,629	4,509	1,195	1,457	547
	lo	all over the city	17	0.85	95,708	25,757	666,006	9,165	2,629	4,509	2,783	1,457	32
4	hi	don't have to worry	407	20.35	179,461	208,841	513,572	2,501	14,806	36,437	1,246	3,050	853
	lo	don't have to wait	40	2	179,461	208,841	513,572	3,408	14,806	36,437	711	3,050	345
5	hi	don't know how much	338	16.9	179,461	489,093	69,598	43,515	60,789	7,768	4,044	3,688	581
	lo	don't know how many	203	10.15	179,461	489,093	69,598	16,745	60,789	7,768	3,665	3,688	360
6	hi	go to the doctor	394	19.7	73,108	513,572	666,006	2,272	16,670	23,134	1,129	3,775	688
	lo	go to the beach	139	6.95	73,108	513,572	666,006	1,672	16,670	23,134	855	3,775	290
7	hi	how do you feel	739	36.95	69,598	162,938	845,026	17,699	3,935	46,315	3,106	2,913	1,384
	lo	how do you do	132	6.6	69,598	162,938	845,026	162,938	3,935	46,315	11,043	2,913	3,546
8	hi	I don't know why	957	47.85	137,715	179,461	489,093	18,560	119,395	60,789	1,743	55,116	1,017
	lo	I don't know who	232	11.6	137,715	179,461	489,093	35,990	119,395	60,789	1,916	55,116	402
9	hi	I have a lot	675	33.75	137,715	208,841	518,226	64,812	32,615	33,649	60,720	8,385	3,447
	lo	I have a little	225	11.25	137,715	208,841	518,226	36,301	32,615	33,649	23,690	8,385	835
10	hi	I have to say	420	21	137,715	208,841	513,572	39,886	32,615	36,437	7,537	3,864	1,166
	lo	I have to see	28	1.4	137,715	208,841	513,572	50,245	32,615	36,437	8,386	3,864	201
11	hi	I want to go	256	12.8	137,715	30,722	513,572	73,108	4,239	18,081	20,031	2,571	1,593
	lo	I want to know	78	3.9	137,715	30,722	513,572	489,093	4,239	18,081	3,558	2,571	442
12	hi	it's kind of hard	342	17.1	248,951	51,585	358,463	14,236	5,292	49,858	660	5,188	571
	lo	it's kind of funny	181	9.05	248,951	51,585	358,463	10,916	5,292	49,858	453	5,188	406
13	hi	on the other hand	734	36.7	131,496	666,006	39,286	3,278	30,450	12,224	766	1,440	752
	lo	on the other end	96	4.8	131,496	666,006	39,286	6,738	30,450	12,224	207	1,440	198
14	hi	out of the house	240	12	74,497	358,463	666,006	9,926	13,508	49,334	2,929	2,890	399
	lo	out of the game	16	0.8	74,497	358,463	666,006	5,958	13,508	49,334	1,270	2,890	195
15	hi	we have to talk	252	12.6	186,179	208,841	513,572	17,366	15,528	36,437	7,018	2,288	493
	lo	we have to say	18	0.9	186,179	208,841	513,572	39,886	15,528	36,437	7,537	2,288	1,166
16	hi	where do you live	1063	53.15	49,152	162,938	845,026	22,382	2,236	46,315	3,624	1,944	1,907
	lo	where do you work	87	4.35	49,152	162,938	845,026	28,315	2,236	46,315	1,541	1,944	572

APPENDIX C: Fillers in Experiment I

Table 17. List of fillers in the phrasal acceptability judgment task

12 possible sequences	68 impossible sequences		
	17 fillers with an incorrect preposition	51 fillers with a wrong word order	
1. hold a green bag	13. put from the shelf	30. six weeks past have	56. at look her watch
2. I kicked the ball	14. look with the sky	31. girl the didn't sleep	57. the at same time
3. center of the stage	15. jump during the pool	32. the in room next	58. a glass wine of
4. picture of the garden	16. was living at himself	33. girl the was sad	59. tea the is sweet
5. the boy was mean	17. proud on myself	34. to dance him with	60. a on white plate
6. my only guess is	18. was talking out Paul	35. for closed two weeks	61. table the was brown
7. the girl won't move	19. left home out Sunday	36. hold it way this	62. changed he his clothes
8. ten weeks are gone	20. afraid to the dark	37. Ted blue had eyes	63. Tim a book wrote
9. on the whiteboard	21. met of Union Street	38. over climb the hill	64. she loud laughed out
10. John had the flu	22. put up to Paul	39. Sue some ate pasta	65. in first the year
11. the salad was great	23. stood next out Jim	40. that radio broken was	66. I my found keys
12. buy a new dress	24. he depended in her	41. computers very are useful	67. to tell truth the
	25. excited above the news	42. cut some him bread	68. look the at screen
	26. met down the morning	43. she legs long has	69. a of group people
	27. arrive about England	44. for of learners English	70. John his met teacher
	28. explain it at Kate	45. hanging the on wall	71. it all snowed week
	29. knock about the door	46. on brown the seat	72. she a has boat
		47. dogs good are pets	73. under tree the
		48. mom my was strict	74. Larry replaced watch his
		49. Chris a made cake	75. Kate the plays piano
		50. bus the late was	76. ocean the is polluted
		51. Bill made cookies some	77. painted Laura it green
		52. walk across hill the	78. she her boyfriend called
		53. the dogs wet got	79. a cup coffee of
		54. a bought book new	80. he emailed boss his
		55. looking for man the	

APPENDIX D: Background questionnaire for native English speakers

Background questionnaire for native English speakers: Part I

Please provide the following information about yourself.

1. Age: _____
2. Gender: ☐ Male ☐ Female
3. Please select the type of your current academic program. If you are not a student, please indicate the highest degree you have received.

   ☐ Bachelor's ☐ Master's ☐ PhD Major: _____
4. Have you ever had ☐ a vision problem, ☐ hearing impairment, ☐ speech production difficulty? (Check all applicable)
5. Are you a right-handed person or a left-handed person?: _____
6. Your native language: _____
7. Please list the languages you know **in order of dominance** (language that you can speak most fluently first):

1.	2.	3.	4.

8. Please list the languages you know **in order of acquisition** (your native language first):

1.	2.	3.	4.

9. In the table below, please estimated level of proficiency in the language your know on a scale of 1-5 (1 = poor, 5 = excellent)

Language	Speaking	Listening	Reading	Writing

10. Have you ever taken a standardized test of these languages? If yes, please indicate the test name, the language tested, the score you received, and the date when you took it:_____

    _____

11. Please list what percentage of the time you currently and on average **listen to** each language you know. (Your percentages should add up to 100%):

List language here				
List percentage here				

12. Please list what percentage of the time you currently and on average **speak** each language you know. (Your percentages should add up to 100%):

List language here				
List percentage here				

13. Please list what percentage of the time you currently and on average **read** each language you know. (Your percentages should add up to 100%):

List language here				
List percentage here				

14. Please list what percentage of the time you currently and on average **write** each language you know. (Your percentages should add up to 100%):

List language here				
List percentage here				

15. If you have ever lived in another country for more than three months, please provide the name of the country and approximate dates of residence: _____
_____
_____

Background questionnaire for native English speakers: Part II

16. What language do you consider your second language?: _____
All the questions below refer to your knowledge of your second language. Write N/A if inapplicable for any reason.
17. Age when you …

began acquiring it	became fluent in it	began reading in this language	became fluent in reading it

18. Please list the number of years and months you spent in each language environment.

	Years	Months
A country where this language is spoken		
A family where this language is spoken		
A school and/ or working environment		

19. On a scale from zero to ten (0 = not at all, 10 = a lot), please select how much the following factors contributed to you learning of your second language:

Interacting with friends		Self learning	
Interacting with family		Watching TV	
Reading		Listening to the radio	
Others (please specify) _____			

20. In your perception, how much of an accent do you have in your second language?:

☐ Not at all ☐ A little ☐ Quite a lot ☐ A lot

21. Please rate how frequently others identify you as a non-native speaker based on your accent in your second language:

☐ Never ☐ Rarely ☐ Sometimes ☐ Often ☐ Always

APPENDIX E: Background questionnaire for ESL learners

Background questionnaire for ESL learners: Part I

Please provide the following information about yourself.

1. Age: _____
2. Gender: ☐ Male ☐ Female
3. Please select the type of your current academic program. If you are not a student, indicate the highest degree you have received.

   ☐ Bachelor's ☐ Master's ☐ PhD Major: _____
4. Have you ever had ☐ a vision problem, ☐ hearing impairment, ☐ speech production difficulty? (Check all applicable)
5. Are you a right-handed person or a left-handed person?: _____
6. Your native language: _____
7. Please list the languages you know **in order of dominance** (language that you can speak most fluently first):

1.	2.	3.	4.

8. Please list the languages you know **in order of acquisition** (your native language first):

1.	2.	3.	4.

9. In the table below, please list any other languages that you have learned and your estimated level of proficiency on a scale of 1-5 (1 = poor, 5 = excellent)

Language	Speaking	Listening	Reading	Writing

10. At what age did you start learning English? _____
11. How did you learn English up to this point? (check all that apply)
    - Mainly through classroom instruction _____
    - Mainly through interacting with people _____
    - A mixture of both _____
    - Other (specify) _____
12. How long have you studied/ lived in the U.S.? _____
13. Before coming to the US, did you ever live in an English speaking country for more than two months? If yes, please indicate where, when, and how long: _____

    _____

    _____
14. If you took the TOEFL or the IELTS before, please indicate your latest score and the year you took it: _____

_____

15. Have you ever taken a standardized test of other languages? If yes, please indicate the test name, the language tested, the score you received, and the date when you took it:_____

_____

_____

Background questionnaire for ESL learners: Part II

Please provide the following information about yourself.

16. Please list what percentage of the time you currently and on average listen to each language you know. (Your percentages should add up to 100%):

List language here				
List percentage here				

17. Please list what percentage of the time you currently and on average **speak** each language you know. (Your percentages should add up to 100%):

List language here				
List percentage here				

18. Please list what percentage of the time you currently and on average **read** each language you know. (Your percentages should add up to 100%):

List language here				
List percentage here				

19. Please list what percentage of the time you currently and on average **write** each language you know. (Your percentages should add up to 100%):

List language here				
List percentage here				

20. In your perception, how much of an accent do you have in English?
☐ Not at all ☐ A little ☐ I am not sure ☐ Quite a lot ☐ A lot

21. Please rate how frequently others identify you as a non-native speaker based on your accent in English:
☐ Never ☐ Rarely ☐ Sometimes ☐ Often ☐ Always

22. How often do you do the following?
    - Read English academic textbooks, papers, or business reports ☐ Never ☐ Rarely ☐ Sometimes ☐ Often ☐ Always
    - Read English novels, magazines, or ☐ Never ☐ Rarely ☐ Sometimes ☐ Often ☐ Always

newspapers

- Write emails, reports, papers, or essays in English ☐ Never ☐ Rarely ☐ Sometimes ☐ Often ☐ Always
- Watch English TV programs or movies, or listen to English songs or radio programs ☐ Never ☐ Rarely ☐ Sometimes ☐ Often ☐ Always
- Speak English to native English speakers ☐ Never ☐ Rarely ☐ Sometimes ☐ Often ☐ Always
- Speak English to non-native English speakers ☐ Never ☐ Rarely ☐ Sometimes ☐ Often ☐ Always

23. How important is it to you to learn English?
☐ not important at all ☐ not really ☐ so-so ☐ quite important ☐ very important

24. How much do you like learning English?
☐ not at all ☐ not really ☐ so-so ☐ quite a lot ☐ very much

25. How much would you like to become similar to native English speakers?
☐ not at all ☐ not really ☐ so-so ☐ quite a lot ☐ very much

APPENDIX F: Transformation of regression coefficients for result interpretations

Because the dependent variables in the current study (reaction time in Experiment I and production durations of the first three words in the target phrases in Experiment II) were on the base-10 logarithmic scale, the regression coefficients were transformed for more meaningful result interpretations. For example, in Experiment I (Table 7), in the case of phrase frequency, which is a binary predictor, the meaning of the regression coefficient β is as follows:

$$\Delta Y / \Delta X = \beta$$

$$(\log RT_1 - \log RT_0) / \Delta X = \beta$$

RT denotes reaction time, and 0 and 1 are associated with the reference category (low frequency) and the non-reference category (high frequency), respectively. Because X = 0 for the reference category and X = 1 for the non-reference category, ΔX (the difference in the X values) equals 1. Moreover, based on Table 7, the regression results indicate that β for frequency condition is -0.04. Therefore:

$$\log RT_1 - \log RT_0 = -0.04$$

$$10^{\log RT_1 - \log RT_0} = 10^{-0.04} \text{ (Take the exponential of both sides)}$$

Since $10^{\log RT_1 - \log RT_0} = 10^{\log RT_1} / 10^{\log RT_0}$ (e.g., $10^{5-3} = 10^5 / 10^3 = 10^2$), the following was obtained:

$$10^{\log RT_1} / 10^{\log RT_0} = 10^{-0.04}$$

Moreover, $10^{\log RT_1}$ equals RT_1 (e.g., If RT_1 = 100 ms, then $10^{\log 100} = 10^2 = 100$) and similarly $10^{\log RT_0}$ equals RT_0. Therefore:

$$RT_1 / RT_0 = 10^{-0.04}$$

$$RT_1 = (10^{-0.04}) * RT_0 = 0.91 * RT_0$$

This means that on average the reaction time for the non-reference category (high frequency) was 10^{β} times or 0.91 times the reaction time for the reference category (low frequency). That is, in this case, the participants were on average 9% faster when judging acceptability of high frequency phrases. The interpretations for continuous predictors (e.g., the number of characters in Table 7) were also along this same line. In that case, ΔX also equals to 1 because in the current study continuous predictors were standardized; consequently, a regression coefficient indicates the change in the dependent variable on the base-10 logarithmic scale associated with a one *SD* change in the standardized numerical predictor (i.e., $\Delta X = 1$).

APPENDIX G: Fillers in Experiment II

Table 18. List of fillers in the elicited production task

1. hold a green bag	29. on the brown seat
2. I kicked the ball	30. dogs are good pets
3. center of the stage	31. my mom was strict
4. picture of the garden	32. Chris made a cake
5. the boy was mean	33. the bus was late
6. my only guess is	34. Bill made some cookies
7. the girl won't move	35. walk across the hill
8. ten weeks are gone	36. the dogs got wet
9. on the whiteboard	37. bought a new book
10. John had the flu	38. looking for the man
11. the salad was great	39. look at her watch
12. buy a new dress	40. at the same time
13. six weeks have past	41. a glass of wine
14. the girl didn't sleep	42. the tea is sweet
15. in the next room	43. on a white plate
16. the girl was sad	44. the table was brown
17. to dance with him	45. he changed his clothes
18. closed for two weeks	46. Tim wrote a book
19. hold it this way	47. she laughed out loud
20. Ted has blue eyes	48. in the first year
21. climb over the hill	49. I found my keys
22. Sue ate some pasta	50. to tell the truth
23. that radio was broken	51. look at the screen
24. computers are useful	52. a group of people
25. cut him some bread	53. John met his teacher
26. she has long nails	54. it snowed all week
27. for learners of English	55. she has a boat
28. hanging on the wall	56. take a look at

REFERENCES

REFERENCES

Abbot–Smith, K., & Tomasello, M. (2006). Exempla schematization in a usage based account of syntactic acquisition. *The Linguistic Review, 23*, 275–290.Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language, 42*, 239–273.

Ambridge, B., & Lieven, E. M. (2011). *Child language acquisition: Contrasting theoretical approaches*. Cambridge: Cambridge University Press.

Arnon, I. (2015). What can frequency effects tell us about the building blocks and mechanisms of language learning? *Journal of Child Language, 42*, 274–277.

Arnon, I., & Priva, U. C. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech, 56*, 349–371.

Arnon, I., & Priva, U. C. (2014). The changing effect of word and multiword frequency on phonetic duration for highly frequent sequences. *The Mental Lexicon, 9*, 377–400.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language, 62*, 67–82.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for the relationship between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech, 47*, 31–56.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. New York: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science, 19*, 241–248.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* (CD–ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania

Barr, D. J., Levy, R., Scheepers, C., & Tilly, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and*

Language, 68, 255–278.

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R*. New York: Springer.

Bates, D. M., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48

Bell, A., Brenier, J., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language, 60*, 92–111.

Bell, A., Jurafsky, D., Fosler–Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America, 113*, 1001–1024.

Bley–Vroman, R. (2009). The evolving context of the Fundamental Difference Hypothesis. *Studies in Second Language Acquisition, 31*, 175–98.

Bod, R., Hay, J., & Jannedy, S. (Eds.) (2003). *Probabilistic linguistics*. Cambridge, MA: MIT Press.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research, 10*, 245–261.

Boers, F., & Lindstromberg, S. (2012). Experimental and intervention studies on formulaic sequences in a second language. *Annual Review of Applied Linguistics, 32*, 83–110.

Boersma, P., & Weenink, D. (2010). *Praat: Doing phonetics by computer*. (Version 6.0.09.) [software] Available at http://www.fon.hum.uva.nl/praat/

Bybee, J. (2006). From usage to grammar: The minds response to repetition. *Language, 82*, 711–733.

Bybee, J. (2008). Usage-based grammar and second language acquisition. In P. Robinson & N.C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp.216–236). New York: Routledge.

Bybee, J. (2010). *Language, usage and cognition.* New York: Cambridge University Press.

Bybee, J., & Hopper, P. (2001). *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins.

Bybee, J., & Scheibman, J. (1999). The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics, 37*, 575–596.

Bybee, J., & Thompson, S. (2000). Three frequency effects in syntax. *Berkeley Linguistic Society, 23*, 65–85.

Casenhiser, D., & Goldberg, A. E. (2005). Fast mapping between a phrasal form and meaning. *Developmental Science, 8*, 500–508.

Cedrus Corporation. (2006). *SuperLab Pro* (4.5) [computer software]. San Pedro, CA.

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science, 26*, 609–651.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review, 113*, 234–272.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Christiansen, M. H., & Chater, N. (Eds.). (2001). *Connectionist psycholinguistics*. Westport, CO: Ablex.

Cieri, C., Miller, D., Walker, K., (2004). The Fisher corpus: A resource for the next generations of speech-to-text. *Proceedings of LREC 2004: Fourth International Conference on Language Resources and Evaluation.*

Cieślicka, A. (2006). Literal salience in on-line processing of idiomatic expressions by second language learners. *Second Language Research, 22*, 115–144.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics, 29*, 72–89.

Croft, W., & Cruse, D. A. (2004). *Cognitive linguistics*. Cambridge, UK: Cambridge University Press.

Cunnings, I., & Finlayson, I. (2015). Mixed effects modeling and longitudinal analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 159–181). New York: Routledge.

Davies, M. (2013). Google scholar and COCA-academic: Two very different approaches to examining academic English. *Journal of English for Academic Purposes, 12*, 155–165.

De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2013). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics, 36*, 223–24.

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22*, 499–533.

den Dikken, M. (1995). *Particles: On the syntax of verb-particle, triadic, and causative constructions*. Oxford: Oxford University Press.

Diependaele, K., Lemhöfer, K., & Brysbaert, M. (2013). The word frequency effect in first- and second-language word recognition: A lexical entrenchment account. *The Quarterly Journal of Experimental Psychology, 66*, 843–863

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology, 25*, 104–123.

Durrant, P., & Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research, 26*, 163–188.

Duyck, W., Vanderelst, D., Desmet, T., & Hartsuiker, R. J. (2008). The frequency-effect in second-language visual word recognition. *Psychonomic Bulletin & Review, 15*, 850–855.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking and points of order. *Studies in Second Language Acquisition, 18*, 91–126.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition, 24*, 143–188.

Ellis, N. C. (2003). Constructions, chunking, and connectionism: The emergence of second language structure. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition*. Oxford: Blackwell.

Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition, 27*, 305–352.

Ellis, N. C. (2006a). Language acquisition as rational contingency learning. *Applied Linguistics, 27*, 1–24.

Ellis, N. C. (2006b). Selective attention and transfer phenomena in L2 acquisition: contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics, 27*, 164–194.

Ellis, N. C. (2008a). The dynamics of second language emergence: Cycles of language use, language change, and language acquisition. *Modern Language Journal, 92*, 232–249.

Ellis, N. C. (2008b). The psycholinguistics of the interaction hypothesis. In A. Mackey and C. Polio (Eds.), *Multiple perspectives on interaction in SLA: Second language research in honor of Susan M. Gass* (pp. 11–40). New York: Routledge.

Ellis, N. C. (2011). Frequency-based accounts of SLA. In S. Gass & A. Mackey (Eds.), *Handbook of second language acquisition*, (pp. 193–210), London: Routledge/Taylor Francis.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics, 32*, 17–44.

Ellis, N. C. (2013). Second language acquisition. In G. Trousdale & T. Hoffmann (Eds.), *Oxford handbook of construction grammar* (pp. 365–378). Oxford: Oxford University Press.

Ellis, N. C. & Cadierno, T. (Eds). (2009). Constructing a second language. *Annual Review of Cognitive Linguistics, Special Section. 7*, 111–290.

Ellis, N. C., & Laporte, N. (1997). Contexts of acquisition: Effects of formal instruction and naturalistic exposure on second language acquisition. In A. M. B. de Groot & J. F. Kroll (Eds.), *Tutorials in bilingualism: Psycholinguistic perspectives* (pp. 53–83). Mahwah, NJ: Erlbaum.

Ellis, N. C., & Larsen–Freeman, D. (2009). Constructing a second language: Analyses and computational simulations of the emergence of linguistic constructions from usage. *Language Learning, 59*, 90–125.

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2013). Usage-based language: Investigating the latent structures that underpin acquisition. *Language Learning, 63*, 25–51.

Ellis, N. C., & Ferreira–Junior, F. (2009a). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics, 7*, 187–220.

Ellis, N. C., & Ferreira–Junior, F. (2009b). Construction learning as a function of frequency, frequency distribution, and function. *Modern Language Journal, 93*, 370–385.

Ellis, N. C., Simpson–Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly, 42*, 375–396.

Ellis, N. C. & Wulff, S. (2015). Second language acquisition. In Dabrowska, E., & Divjak, D. (Eds.), *Handbook of cognitive linguistics* (pp. 409–431). Berlin, Germany: De Gruyter Mouton.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff–Smith, A., Parisi, D., & Plunkett,K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Eskildsen, S. W. (2012). L2 negation constructions at work. *Language Learning, 62*, 335–372.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp.1212–1248). Berlin: Mouton de Gruyter.

Faraway, J. J. (2006). *Extending the linear model with R*. Boca Raton, FL: Chapman & Hall/CRC.

Feldweg, H. (1991). *The European science foundation second language database.* Nijmegen, The Netherlands: Max Planck Institute for Psycholinguistics. Available at http://www.mpi.nl/world/tg/lapp/esf/esf.html.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London, UK: SAGE Publications.

Fox, J. (2008). *Applied regression analysis and generalized linear models* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Fulga, A., & McDonough, K. (2016). The impact of first language background and visual information on the effectiveness of low-variability input. *Applied Psycholinguistics, 37*, 265–283.

Godfrey, J. J., Holliman, E. C. & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 1*, 517–520.

Godfroid, A., & Uggen, M. S. (2013). Attention to irregular verbs by beginning learners of German. *Studies in Second Language Acquisition, 35*, 291–322.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure.* Oxford, UK: Oxford University Press.

Goldberg, A. E. (1999). The emergence of the semantics of argument structure constructions. In B. MacWhinney (Ed.), *Emergence of language*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences, 7*, 219–224.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Goldberg, A. E. (2013). Argument structure constructions versus lexical rules or derivational verb templates. *Mind & Language, 28*, 435–465.

Goldberg, A. E., & Casenhiser, D. (2008). Construction learning and second language acquisition. In P. Robinson & N.C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp.197–215). New York: Routledge.

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics, 15*, 289–316.

Goldberg, A. E., Casenhiser, D. M., & White, T. (2007). Constructions as categories of language. *New Ideas in Psychology, 25*, 70–86.

Gollan, T. H., Montoya, R. I., Cera, C., & Sandoval, T. C. (2008). More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language, 58*, 787–814.

Gollan, T. H., Montoya, R. I., & Werner, G. (2002). Semantic and letter fluency in Spanish–English bilinguals. *Neuropsychology, 16*, 562-57

Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab–Aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics. 39*, 192–193.

Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly, 25*, 373–406.

Gries, S. T. (2010). Useful statistics for corpus linguistics. In A. Sánchez & M. Almela (Eds.), *A mosaic of corpus linguistics: Selected approaches* (pp. 269–291). Frankfurt, Germany: Peter Lang.

Gries, S. T. (2015). 50-something years of work on collocations. *International Journal of Corpus Linguistics, 18*, 137–165.

Gries, S. T., & Berez, A. L. (2016) Linguistic annotation in/for corpus linguistics. In Nancy I. & James P. (Eds.), *Handbook of linguistic annotation*. Berlin/ New York: Springer.

Gries, S. Th., & Ellis, N.C. (2015). Statistical measures for usage-based linguistics. *Language Learning, 2,* 228–255.

Gyllstad, H., & Wolter, B. (2016). Collocational processing in light of the phraseological continuum model: Does semantic transparency matter? *Language Learning, 66,* 296–323.

Hernández, M., Costa, A., & Arnon, I. (2016). More than words: Multiword frequency effects in non-native speakers. *Language, Cognition and Neuroscience, 31,* 785–800.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.

Hoffmann, T., & Trousdale, G., (Eds.), *Oxford handbook of construction grammar*. Oxford: Oxford University Press.

Hout, A.V.D., Fox, J. –P., & Muniz–Terrera, G. (2015). Longitudinal mixed-effects models for latent cognitive function. *Statistical Modelling, 15*, 366–387.

Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing* (2nd ed.). Ann Arbor, MI: University of Michigan Press.

Ibbotson P. (2013). The scope of usage-based theory. *Frontier in Psychology, 4*. doi: 10.3389/fpsyg.2013.00255

Jannsen, N., & Barber, H. A. (2012). Phrase frequency effects in production. *PLoS ONE, 7. Retrieved from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0033202*

Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics, 21*, 47–77.

Jiang, N., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *Modern Language Journal, 91*, 433–445.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science, 20*, 137–194.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. (2000). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins.

Khang, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: temporal measures and stimulated recall. *Language Learning, 64*, 809–854.

Kim, S.–H., & Kim, J.–H. . (2012). Frequency effects in L2 multiword unit processing: Evidence from self-paced reading. *TESOL Quarterly, 46*, 831–841.

Langacker, R. W. (2008). *Cognitive grammar: A basic introduction.* New York: Oxford University Press.

Larsen–Freeman, D., & Long, M. (1991). *An introduction to second language acquisition research*. New York: Longman.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61*, 647–672.

Leech, G. (1992). 100 million words of English: The British National Corpus. *Language Research, 28*, 1–13.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods, 44*, 325–343.

Littlemore, J. (2009). *Applying cognitive linguistics to second language learning and teaching*. Basingstoke, UK: Palgrave Macmillan.

Loewen, S., & Plonsky, L. (2016). *An A–Z of applied linguistics research methods*. New York: Palgrave Macmillan.

MacWhinney, B. (1992). Transfer and competition in second language learning. In R. J. Harris (Ed.), *Cognitive processing in bilinguals* (pp. 371–390). Amsterdam, The Netherlands: North Holland.

MacWhinney, B. (1997). Second language acquisition and the competition model. In A. M. B. De Groot & J. F. Kroll (Eds.), *Tutorials in bilingualism: Psycholinguistic perspectives* (pp. 113–142). Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2008). A unified model. In P. Robinson & N.C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp.341–371). New York: Routledge.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP–Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research, 50*, 940–967.

Matthews, D., Lieven, E., Theakston, A., & Tomasello, M. (2005). The role of frequency in the acquisition of English word order. *Cognitive Development, 20*, 121–136.

McDonald, S. A., & Shillcock, R. (2003). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research, 43*, 1735–1751.

McDonough, K., & Nekrasova–Becker, T. (2014). Comparing the effect of skewed and balanced input on English as a foreign language learners' comprehension of the double-object dative construction. *Applied Psycholinguistics, 35*, 419–442.

McDonough, K., & Trofimovich, P. (2013). Learning a novel pattern through balanced and skewed input. *Bilingualism: Language and Cognition, 16*, 654–662.

McIntyre, A. (2001). *German double particles as preverbs: Morphology and conceptual semantics*. Tübingen, Germany: Stauffenburg.

Meara, P. (2005). *Y_lex* [computer software]. Swansea, UK: Lognostics.

Meara, P., & Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grunwell (Ed.), *Applied linguistics in society: British studies in applied linguistics 3* (pp. 80–87). London: CILT.

Meara, P., & Jones, G. (1990). *Eurocentres vocabulary size tests 10KA*. Zurich, Switzerland: Eurocentres Learning Service.

Muñoz, C. (2008). Symmetries and asymmetries of age effects in naturalistic and instructed L2 learning. *Applied Linguistics, 29*, 578–596.

Myers, R. (1990). *Classical and modern regression with applications* (2nd ed.). Boston, MA: Duxbury.

Nakamura, D. (2012). Input skewedness, consistency, and order of frequent verbs in frequency-driven second language construction learning: A replication and extension of Casenhiser and Goldberg (2005) to adult second language acquisition. *International Review of Applied Linguistics, 50*, 31–67.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam, the Netherlands: Benjamins.

Nippold, M. A., & Rudzinski, M. (1993). Familiarity and transparency in idiom explanation: A developmental study of children and adolescents. *Journal of Speech and Hearing Research, 36*, 728–737.

Ortega, L. (2013). SLA for the 21st century: Disciplinary progress, transdisciplinary relevance, and the bi/multilingual turn. *Language Learning, 63*, 1–24

Ortega, L., Tyler, A. E., Park, H. E., & Uno, M. (2016). *The usage-based study of language learning and multilingualism*. Georgetown, Washington D.C.: Georgetown University Press.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Richards, J. C. and Schmidt, R. W. (Eds.), *Language and communication* (pp.191–225). London, UK: Longman.

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods, 162*, 8–13.

Pinker, S. (1994). *The language instinct*. New York: William Morrow.

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York: Basic Books.

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

Pitt, M., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler–Lussier, E. (2007). *Buckeye corpus of conversational speech* (2nd release). Available from http://buckeyecorpus.osu.edu/.

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

Prasada, S., Pinker, S., & Snyder, W. (1990, November). *Some evidence that irregular forms are retrieved from memory but regular forms are rule-governed*. Paper presented at the 31st meeting of the Psychonomic Society, New Orleans: CA.

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Reali, F., & Christiansen, M. H. (2007). Word-chunk frequencies affect the processing of pronominal object-relative clauses. *Quarterly Journal of Experimental Psychology, 60*, 161–170.

Reddy, S., & Stanford, J. (2015). Toward completely automated vowel extraction: Introducing DARLA. *Linguistics Vanguard, 1,* 15–28.

Robinson, P. & Ellis, N. C. (2008). *Handbook of cognitive linguistics and second language acquisition*. New York: Routledge.

Römer, U., O'Donnell, M. B., & Ellis, N. C. (2014). Second language learner knowledge of verb–argument constructions: Effects of language transfer and typology. *Modern Language Journal, 98*, 952–975.

Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H., Yuan, J. (2014). *FAVE (Forced Alignment and Vowel Extraction)*. Program Suite v1.2.2 10.5281/zenodo.22281.

Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition (Vol. 2: Psychological and biological models)*. Cambridge, MA: MIT Press.

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing, and use* (pp. 127–151). Philadelphia: John Benjamins.

Schmitt, N., & Underwood, G. (2004). Exploring the processing of formulaic sequences through a self-paced reading task. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing, and use* (pp. 173–189). Philadelphia: John Benjamins.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Siyanova–Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research, 27*, 251–272.

Siyanova–Chanturia, A., Conklin, K., & van Heuven, W. J. B. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology, 37*, 776–784.

Siyanova–Chanturia, A., & Martinez, R. (2014). The idiom principle revisited. *Applied Linguistics*, *36*, 549–569.

Sonbul, S. (2015). Fatal mistake, awful mistake, or extreme mistake? Frequency effects on off-line/on-line collocational processing. *Bilingualism: Language and Cognition, 18*, 419–437.

Sosa, A. V., & MacFarlane, J. (2002). Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word *of*. *Brain and Language, 83*, 227–236.

Stengers, H., Boers, F., Housen, A., & Eyckmans, J. (2011). Formulaic sequences and L2 oral proficiency: Does the type of target language influence the association? *International Review of Applied Linguistics in Language Teaching, 49*, 321–343.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Tomasello, M. (2009). The usage-based theory of language acquisition. In E. Bavin (ed.), *Handbook of Child Language*, pp. 69–87. New York: Cambridge University Press.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning, 61*, 569–613.

Tremblay, A., & Tucker, B. V. (2011). The effects of n-gram probabilistic measures on the recognition and production of four-word sequences. *Mental Lexicon, 6*, 302–324.

Tyler, A. (2010). Usage-based approaches to language and their applications to second language learning. *Annual Review of Applied Linguistics, 30*, 270–291.

Ullman, M. T. (1999). Acceptability ratings of regular and irregular past tense forms: evidence for a dual-system model of language from word frequency and phonological neighbourhood effects. *Language and Cognitive Processes, 14*, 47–67.

Ullman, M. T. (2001). A neurocognitive perspective on language: The declarative/procedural model. *Nature Reviews Neuroscience, 2*, 717–726.

Ullman, M. T., Pancheva, R., Love, T., Yee, E., Swinney, D., & Hickok, G. (2005). Neural correlates of lexicon and grammar: Evidence from the production, reading, and judgment of inflection in aphasia. *Brain and Language, 93*, 185–238.

Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences* (pp. 153–172). Amsterdam, the Netherlands: John Benjamins.

Valsecchi, M., Künstler, V., Saage, S., White, B. J., Mukherjee, J., & Gegenfurtner, K. R. (2013). Advantage in reading lexical bundles is reduced in non-native speakers. *Journal of Eye Movement Research, 6*, 1–16.

Webb, S., Newton, J., & Chang, A. (2013). Incidental learning of collocation. *Language Learning, 63*, 91–120.

Whitford, V., & Titone, D. (2012). Second-language experience modulates first and second-language word frequency effects: Evidence from eye movement measures of natural paragraph reading. *Psychonomic Bulletin & Review, 19*, 73–80.

Winter, B. (2013). *Linear models and linear mixed effects models in R with linguistic applications*. Retrieved from https://arxiv.org/ftp/arxiv/papers/1308/1308.5499.pdf

Wolter, B., & Gyllstad, H. (2013). Frequency of input and L2 collocational processing. *Studies in Second Language Acquisition, 35*, 451–482.

Wood, D. (2006). Uses and functions of formulaic sequences in second language speech: An exploration of the foundations of fluency. *Canadian Modern Language Review, 63*, 13–33.

Wood, D. (2010). *Formulaic language and second language speech fluency: Background, evidence and classroom applications*. New York: Continuum.

Wulff, S., Ellis, N. C., Römer, U. T. E., Bardovi–Harlig, K., & Leblanc, C. J. (2009). The acquisition of tense-aspect: Converging evidence from corpora and telicity ratings. *Modern Language Journal, 93,* 354–369.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge, UK: Cambridge University Press.

Yamashita, J., & Jiang, N. (2010). L1 influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations. *TESOL Quarterly, 44*, 647–668.

Year, J., & Gordon, P. (2009). Korean speakers' acquisition of the English ditransitive construction: The role of verb prototype, input distribution, and frequency. *Modern Language Journal, 93*, 399–417.

Zipf, G. K. (1935). *The psychobiology of language: An introduction to dynamic philology*. Cambridge, MA: MIT Press.