Summer 9-18-2014

Explorations into the Psycholinguistic Validity of Extended Collocations

J. Arianna Morgan
Portland State University

Recommended Citation

Morgan, J. Arianna, "Explorations into the Psycholinguistic Validity of Extended Collocations" (2014). *Dissertations and Theses.* Paper 2005.
https://doi.org/10.15760/etd.2004

Explorations into the Psycholinguistic

Validity of Extended Collocations

by

J. Arianna Morgan

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Arts
in
Teaching English to Speakers of Other Languages

Thesis Committee:
Lynn M. Santelmann, Chair
Susan Conrad
Joel S. Steele

Portland State University
2014

ABSTRACT

This study tests the hypothesis that frequency and collocational association make independent contributions to the processing time of English collocations for L1 and L2 English speakers. The results suggest that these constructs do play a role in the processing of 4-word, corpus-extracted phrases. In this sample, L1 speakers demonstrated reduced processing time for both highly frequent and highly associated phrases, while L2 speakers demonstrated reduced processing time for highly frequent phrases. Evidence exists in the data that highly proficient L2 speakers may develop similar patterns of reduced processing time as L1 speakers. Additionally, some L1 speakers did not show the sensitive to higher levels of association typical of this group. Understanding these contributions has the potential to elucidate the most useful targets of phrasal instruction for ESOL students and the psychological mechanisms of associative learning.

DEDICATION

For my grandparents, Sidney M. & Caroline J. Cleveland and Donald C. J. & Eunice M. McElligott. Wish you could all be here.

ACKNOWLEDGMENTS

The process of conducting this experiment and then writing a thesis has been long and challenging. Since beginning this process over 2 years ago, I have journeyed into the realms of psycholinguistics, quantitative psychology, and computer programming in order to develop a knowledge base that would do justice to this topic. It is a bittersweet moment to be finishing this process because of how much I love this area of research: I don't want to be done, but I am also relieved to have found new, related questions that will keep me entertained for years! Before beginning the report of my research findings, I would like to express my appreciation for those who helped me take advantage of the opportunity for rigorous scholarship.

I would like to extend my thanks to the James R. Nattinger Foundation for the honor of holding the Nattinger Fellowship for 2013-2014. It has been a humbling and wonderful experience because I have been able to, in some small sense, continue Dr. Nattinger's work on formulaic language. This extra funding has really allowed me to put the "scholar" in scholarship, as well as continue to be a representative of the LGBT community in academia.

This thesis would not be complete without extending my gratitude my advisor Lynn Santelmann, who has helped me leave the safety of the trees to a better view of the whole forest. Our discussions on knitting, the history of the Germanic languages, psychology, neurology, nativism versus emergentism, religion, and society helped keep me going through the struggles I had with writing, while her

amused, and mildly exasperated, comments on my drafts kept me from digressing into my latest flight of research fancy—or employing the English subjective in strange places.

Thanks also to my second reader Susan Conrad who suggested that I attempt to attempt this particular replication study. The day after my class in Corpus Linguistics that Susan asked me, "Why don't you just replicate this study?" changed my life in ways that I couldn't have imagined then. Susan's expertise in writing has encourage me to avoid undue complexity in composition and value clarity of expression more than the endless clausal subordination of my undergraduate years.

I would also like to thank Joel Steele in Portland State's Department of Psychology for helping a lost linguist find some complex roots in quantitative analysis. Joel's guidance in many sessions of office hours has helped me to develop a far deeper understanding of statistics, methodology, and mathematics than I ever thought I would. My world is richer because of how these techniques allow me to better understand the world.

Special thanks goes to Doug Flahive, Norbert Schmitt and Alison Wray for their extremely helpful feedback at AAAL 2014 in Portland, OR, as well as to Nick Ellis for taking the time to answer my questions about his research in email.

I also extend my thanks to my office mates Andrew Dieckhoff and Jennifer Sacklin who graciously agreed to pilot my experimental procedure and put up with my many rambles about linguistics, statistics, and politics. Thanks also to Kimberly Brown, Eric Dodson, Meghan Oswald, and Jennifer Sacklin for their excellent comments about my practice presentation for AAAL. Their comments helped me to focus on and share the most important parts of my research.

My gratitude goes out to Max Parmer, my dearest friend and confidant. Without his continual support through the journey of computer programming, completing this thesis in the way I wanted to—with (mostly) Free and Open Source software—would have likely taken a backseat to convenience.

Finally, I would like to thank Loren Stearman, a close friend and colleague about to journey to the University of New Mexico, for our many discussions of statistics, R, and programming. Thanks also to as Ernesto Aguilar, Courtney Hearon, Jo-Anne Hutter, Julie Nelson, and Margo Russell for our conversations about linguistics and pedagogy.

Without all of your help, guidance, and mentoring, graduate school would have been much more difficult. Although I am a bit sad to be leaving the Department of Applied Linguistics and all of the wonderful work happening here, I look forward to new opportunities for future scholarship.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

In the past 50 years, the work of Noam Chomsky has been especially influential in the study of language. Chomsky pointed out many of the problems underlying Behaviorist accounts of language learning (e.g., the poverty of the stimulus argument) and in response, developed a theory of language wherein innate syntactic structures, expressed in the micro-structure of the brain, were central to a language user's underlying linguistic competence (Elman, Bates, Johnson, Karmiloff-Smith, Parisi & Plunkett, 1996). Chomksy's goal was to model the sentence-generative, combinatorial rules that he hypothesized language users have in their minds and brains. One of the early criticisms of this theory (Chafe, 1968) was that these syntactically-focused theories did not account well for the presence of idioms—phrases for which the meaning is not immediately derivable from their constituent words.

However, in the past 30 years, a set of competing theories have considered language from a different perspective that attempts to address concerns such as that of Chafe. These Emergentist theories consider much of language to be an emergent phenomenon of human neurocognitive apparatus rather than a set of grammatical classes and rules prespecified in the neurological hardware. Some of these theories have placed types of language similar to idioms in more central theoretical roles (e.g., Ellis, 2002; Sinclair, 1991; Wray, 1992, 2002). Such examples of these types of language are to some degree *formulaic*—i.e., they are thought to be stored and retrieved whole from memory.

My interest in formulaic language arose from becoming familiar with Ellis' (2002) review of frequency effects on language acquisition, while simultaneously learning about the tools corpus linguists use to conduct empirical investigations of lexicogrammatical structure. I read the results of studies conducted by Ellis and his research group and became fascinated with the idea that a corpus has statistical properties that seem to predict faster linguistic processing (Ellis & Simpson-Vlach, 2009; Ellis et al., 2008). At first I did not understand the significance of investigating the psycholinguistic validity of these corpus statistics; however, as I began to understand the basis of associational strength of phrases, it became clear that these strengths reflect not so much the state of the corpus, but rather the state of the collective linguistic consciousness of the community(ies) whence the texts were drawn (Ellis, 2002; Langacker, 2000; Wray, 2002). These concepts struck me as especially relevant for understanding language because of the pervasive formulaicity that empirical research has described—namely, that much of language consists of recurrent multiword phrases (Biber, Johannson, Leech, Conrad & Finegan, 1999; Erman & Warren, 2000; Sinclair, 1991).

Interest in formulaic language has grown substantially in recent years across a variety of fields in applied linguistics (Wray, 2012), including clinical linguistics (Bridges & Van Lancker Sidtis, 2013; Lindholm & Wray, 2011; Sidtis, Canterucci & Katsnelson, 2009; Van Lancker Sidtis & Postman, 2006; Wray, 2010) ESL, EAP and ESP (Ädel & Erman, 2012; Biber, 2006; Biber, Conrad & Cortes, 2004; Durrant & Schmitt, 2010; Simpson-Vlach & Ellis, 2010; Wray, 2000), FLA (Arnon & Snider, 2010; Bannard & Matthews, 2008; Ellis & Simpson-Vlach, 2009), and SLA (Conklin & Schmitt, 2008; Durrant & Doherty, 2010; Ellis, Simpson-Vlach & Maynard, 2008; Eskildsen, 2009,0; O'Donnell, Römer & Ellis, 2013). Thus, this study comes at a time with many new findings, yet few attempts to replicate previous research. This apparent lack of replication also contributed to my desire to undertake a partial replication of Ellis and colleagues' work on the processing of *n*-grams, which are any sequence of graphemes, phonemes, syllables, or lexemes that can be extracted from a corpus (Ellis & Simpson-Vlach, 2009; Ellis, Simpson-Vlach & Maynard, 2008). Since the time of publication of these particular foundational studies until the present, no researchers in an independent lab have attempted to replicate the findings on the processing of *n*-grams (O'Donnell et al., 2013).

Given this opportunity for scholarship, I elected to conduct an experimental study to investigate the effects of frequency and collocational association on the this processing of formulaic language by L1 and L2 speakers of English. My study involves nested, multilevel measurements: Each stimulus item has two independent variables of interest associated with it. These variables are degree of association and frequency. Similarly, each participant is either part of the L1 or L2 speaker groups. This design is methodologically similar to Ellis et al.'s; however, I have chosen to employ a multivariate statistical technique that explicitly factors in the multilevel nature of the data.

Over the next four chapters of this thesis, I shall make the case that certain types of frequency, highly associated recurrent sequences do, in fact, impart a reduction in processing time. Further, I will argue that these findings can inform ESOL teachers and materials developers decisions about which sorts phrases may be targets for instruction. The results of this study also contribute further to understanding processing of one type of formulaic language.

In Chapter 2, I shall review three bodies of literature. First, I shall examine in more detail the nature of formulaicity in language as well as the terminology used to describe it. In doing so, I will show that formulaic language may take many forms with different degrees of semantic transparency, frequency, and degrees of association. Second, I shall discuss how textual corpora can be used in the automatic identification of some types of formulaic language. Finally, I will review studies from First Language Acquisition (FLA) and Second Language Acquisition

(SLA). These studies will examine a) the use of corpus-identified sequences by first and second language speakers of English and b) whether a processing advantage associated with corpus-identified formulaic language exists.

In Chapter 3, I will present the methodology of the study. Of primary concern in this section are the research design, participants, the identification of experimental stimuli, the data collection process, and the description of the variables of interest. Additionally, I shall justify the decision to employ a mixed-effects model, which represents a departure from the statistical analyses that Ellis and colleagues employed.

In Chapter 4, I will then report the results of this partial replication study. The results, although not a perfect replication of Ellis et al.'s (2008), do suggest that frequent, highly associated multiword sequences are, in fact, psychologically real, resulting in reduced processing time. I will also show that there is evidence that L2 speakers can develop similar patterns of reduced processing time that L1 speakers have.

In Chapter 5, I shall discuss the implications and directions for future research. I describe a thought experiment exploring the practical significance of the results from Chapter 4. Further discussion will include considering the implications of this research for the contexts of TESOL and psycholinguistics and potential directions for future research in the development of associational measures for corpora and neurological studies of language pathology and development. These implications and directions for future research may provide valuable avenues of inquiry for language education professionals, psycholinguists, quantitative linguists, and neurolinguists.

Chapter 2

LITERATURE REVIEW

In this chapter, I shall review three bodies of literature relevant to my study. First, I shall consider the nature of formulaicity in language. Second, I will discuss the profusion of terminology within the field of formulaic language studies in order to highlight formulaic language types particularly relevant for my study. I will also show that formulaic language may take many forms with different degrees of semantic transparency, frequency, and degrees of association. Third, I shall discuss how textual corpora can be used in the automatic identification of some types of formulaic language. I will then connect theories of language processing and production that have arisen from corpus-based research to emergentist models of language. Finally, I will review studies from First Language Acquisition (FLA) and Second Language Acquisition (SLA). These studies will examine a) the use of corpus-identified sequences by first and second language speakers of English and b) whether a processing advantage associated with corpus-identified formulaic language for L1 or L2 speakers of English exists.

2.1 FORMULAIC LANGUAGE

Formulaic language (FL) is a general term for many different types of fully and partially fixed linguistic sequences that are thought to be stored and retrieved as whole units from memory, rather than being generated each time a speaker wishes to convey a particular message (Ellis, 2002; Wray, 2002; Wray & Perkins, 2000). FL is theorized to aid in the processing and production of language. For instance, Pawley & Syder (1983) hypothesized that the use of FL reduces the cognitive load of processing and producing language. They observed that although language users are able to generate an infinite number of novel utterances, they do not constantly employ this faculty. Similarly Wray (2002) has argued that FL helps to decrease the cognitive load for processing and production because storing and retrieving multiword sequences from memory requires fewer linguistic pieces to be put together and thus puts less strain on working memory. The freed cognitive resources can then be directed towards making meaning to achieve the user's social goals and needs (Wray, 2002).

Under the theory put forth by Pawley & Syder and Wray, a phrase like *on the other hand*, which has a very particular discursive function (i.e. contrast) and is idiomatic, frequent, and has a high degree of association between its constituent words, is likely retrieved as a whole unit rather than being generated as a Prepositional Phrase (PP) with a Noun Phrase (NP) as the complement of the preposition, with single lexemes filling in the appropriate grammatical slots. However, such theories do not preclude the presence of a concurrent analytic language production and processing mechanism. For instance, Wray (2002) also theorized that the lexicon is heteromorphic—namely, that a multiword sequence may be stored holistically, while simultaneously containing separate representations of the constituent parts on the lexical and morphophonological levels. These constituent parts have been derived through the unconscious analysis of holistic chunks in a child-learner's input (Wray, 1992, 2002). This multiple representation allows for the analysis and decomposition of these larger sequences. Wray's argument is not that formulaic language cannot be analyzed by the language grammar, but that a human's default processing mode is for longer sequences, so as to allow for greater attentional focus on the decoding and juxtaposition of propositions (Wray, 1992). Analysis is thus invoked only on an as-needed-basis.

Ellis (2002), operating from a similar position, though more influenced by connectionist modeling strategies (e.g., Rumelhart & McClelland, 1986), has argued that when words occur in proximity to each other, they become more closely associated with one another: Fluent use of language reflects the ability to recall these chunks by making use of the implicit sense of lexical co-occurrence that humans are thought possess. The process Ellis described is similar to Wray's account, in that both perspectives allow analytic capabilities to develop through the abstraction of schemata from larger sequences of input. Since humans seem able to hold about 4 to 7 chunks of information in working memory, storing multiword, formulaic chunks would allow for more rapid processing and production of language than constant, on-line generation of all clausal elements (Ellis, 2001).

With this brief introduction to FL in mind, I will discuss several types of more easily observable types of FL that will provide the framework for my study, which will use n-grams as corpus-derived psycholinguistic test items. This rest of this section is intended to serve as a very brief introduction to a few FL types—i.e., poems, prayers, idioms, collocations, and n-grams. Each of these types of FL will be treated in more detail in subsequent sections of this literature review.

The first type of FL consists of poems, prayers, songs, plays and nursery rhymes. These pieces of language may persist in one form or another for centuries (cf. Shakespearean plays, the Lord's Prayer). Often these memorized pieces retain vestiges of once common, archaic forms (e.g., A rose by any other name, Our Father who *art* in heaven, *hallowed be thy name*; if this *be* treason) that persist well after older forms have fallen out of common usage. Such preserved, archaic forms presented in contemporary language may then become important objects of evidence in historical linguistics.

Second, idioms are are present in all languages. Idioms at the most basic level are multiword strings, the meaning of which is not derivable from the constituent words of the phrase (Swinney & Cutler, 1979). Common examples of idioms would be: *a piece of cake* (an easy task) or *to have a monkey on one's back* (to have an addiction). However, idioms are not a monolithic category and closer examination reveals several interesting properties of idioms (e.g., varying degrees of fixedness and semantic opacity).

Third, the term collocation was coined by Firth. He is perhaps best remembered for his oft quoted, "You shall know a word by the company it keeps" (Firth, 1957, cited in Wray, 2002). What he meant by this is that a word is not possible to define in isolation but rather that its meaning depends on the contexts in which it occurs. When two words co-occur more often than chance alone would predict, they are termed collocates of each other. However, this designation depends on the words' frequencies, so lacking a computer, it is difficult for humans to accurately gauge what words are collocates. In order to illustrate what a collocation is from a quantitative perspective, I will borrow an example from Durrant & Doherty (2010, p. 130) which examines the occurrence of *strong tea* in the British National Corpus (BNC). This phrase is a collocation because in the 100 million word BNC, *strong* occurs 0.00016% of the time, while *tea* occurs 0.00008% of the time. Based on this, one would expect the sequences *strong tea* and *tea strong* to occur approximately 1.26 times. However, *strong tea* occurs 28 times, which seems very high in comparison to the expected frequency. There are ways to quantify this relation between words (e.g. mutual information score, log-likelihood, *t*-score (Church & Hanks, 1990; Dunning, 1993; Durrant & Doherty, 2010)).

Fourth, the idea of a collocation can be expanded to include multiple words. Lexical bundles arise out of Firth's conception of collocations. Biber et al. (1999) operationalized the term lexical bundles term as "the combinations of words that in fact recur most commonly in a given register" (p. 992) and considered them to be a form of extended collocations. Such sequences are at least three words long and do not have any requirements as far as idiomaticity or structure. Examples of this phenomenon would be *on the other hand* or it has been shown that. These are phrases that occur more often than chance would predict given the frequencies of their constituent words. These phrases, like collocations, can be described both in terms of their absolute frequency in a corpus, as well as how much more or less often than chance they occur.

Finally, as noted above in this section, n-grams are the focus of this study. Although these sequences may be composed of phonemes, graphemes, words, or series of words, I will limit my consideration in this study to those composed of words. Examples of n-grams include but are not limited to: *bird* (sequence of a single word), *caught a fish* (a sequence of words), or *on the other hand* (a sequence of words with an idiomatic meaning). That a sequence is attested in a corpus qualifies it as an n-gram without any requirements for idiomaticity or syntactic constituency—as such, n-grams may be types of non-formulaic language such as neologisms attested only once, non-collocational phrases, or very infrequent idioms. This very broad term can encompass idioms, collocations, and lexical bundles, which are types of formulaic language discussed below. All idioms, collocations, and lexical bundles that are able to be extracted by a computer are n-grams; however, not all n-grams can be so reciprocally classified, as each of these other FL types has more complex criteria for inclusion in the category than simple attestation. N-grams are easy to identify automatically with corpus analytic tools that allow for minimum frequency and associational selection criteria.

In summary, the previous paragraphs have illustrated some properties of several types of FL. However, over the past 30 years, there have been many operationalizations of FL. In the following section, I will consider the terminological profusion that pervades the field in order to better contextualize n-grams within a larger area of research.

2.1.1 Terminological Profusion: Overlap and Divergence

The previous section detailed only a few types of formulaic sequences. Wray & Perkins (2000, p. 3) noted "[w]ell over 40 different terms" that have been applied to different types of formulaic language, including but not limited to: *multiword expressions* (Eskildsen, 2009,0), *multiword sequences* Bannard & Matthews (2008); Arnon & Snider (2010), *formulaic sequences* (Wray & Perkins, 2000; Wray, 2002), *n*-grams (Ellis & Simpson-Vlach, 2009; Ellis et al., 2008), *lexical bundles* (Biber et al., 1999), *prefabs* (Erman & Warren, 2000), *lexical phrases* (Nattinger & DeCarrico, 1992), and *morpheme equivalent units* (Wray, 2008). Although the field may be as Eskildsen (2009, p. 337) observed, "terminologically handicapped" because of this profusion of simultaneously distinct but overlapping terms, not all scholars agree with this perspective. For instance, Wray (2008) noted that although the abundance of terminology might seem unnecessary, it is sometimes useful for researchers to consider different aspects of formulaicity, contingent upon the scope and purpose of their research. Given the terminology breadth of the field, I will explicate some of the most relevant overlaps and divergences to better contextualize the field and this study's place therein.

First, perhaps the broadest of all these terms is that of *formulaic language* (FL). All other terms may be subsumed under this umbrella term. For my purposes, this term encompasses both fully or partially fixed sequences employed by one speaker but also ritualized exchanges that include multiple interlocutors (e.g., Priest: *The Lord be with You*, Laity: *And also with you(r spirit)* or *How're you? Just fine.*).

A *formulaic sequence* is a general term that Wray & Perkins (2000) proposed that encompasses many different types of fully or partially fixed expressions that have been described in the literature. Many other terms can be categorized under this term (i.e., poems, prayers, songs lyrics, idioms, collocations, lexical bundles, and *n*-grams). Employing this generalized definition provides a beginning framework for the identification of a large portion of FL in the wild (Wray, 2008).

Under these broad terms, other distinctions relevant to this study exist as well. For instance, although collocates and lexical bundles are similar in concept, a word's collocates are not necessarily neighboring words. When one examines a word like *deal* one will find that *great* is a collocate to the immediate left of *deal* and that *a* is a collocate to the second left slot of deal. Both of these words are collocates of *deal*, but only one is a contiguous collocate. In contrast, *a great deal*, depending on its frequency across a given register, could be considered a lexical bundle. This lexical bundle could have its own collocations, but the phrase itself, as a lexical bundle, is a single unit.

A particularly relevant example of overlap exists between Eskildsen's multiword-expressions, Bannard & Matthews's multiword sequences, and Biber et al.'s lexical bundles. Although these terms refer to fixed expressions, the primary difference is one of analytic scope. The former two terms were used to describe recurrent sequences in a single person's speech over an extended period of time, whereas the latter describes the most frequent, recurrent sequences, used by multiple speakers or authors, in an entire register. In order to analyze an entire register, the researcher must have available a range of texts, text types, and authors. Thus, lexical bundles are phenomena that exist in the collective consciousness of a linguistic community. These distinctions also illustrate that formulaicity can be something present in both the individual and in a linguistic community.

Another interesting case of overlap comes into play between lexical bundles and idioms. Idioms are marked by some degree of semantic opacity (discussed in more detail below), while lexical bundles are defined by their frequency without any such semantic criteria. Thus idioms, which are not defined by frequency in the same way, can be lexical bundles provided that they are frequent enough. An example of this would *on the other hand*. It is highly associated and quite frequent, occurring in the Corpus of Contemporary English (COCA) at a rate of 42.61 per million in all registers and 83.87 in academic writing (Davies, 2008).

This phrase is used, particularly in academic writing, to invoke contrast between differing positions. It is not semantically transparent because there are no literal hands within a research article (one could argue, though, that neither is it totally semantically opaque either, because the invocation of changing the focus of hands highlights a shift in rhetorical position).

The final highly relevant distinction to make is between lexical bundles and *n*-grams. While both are extracted from large bodies of texts, lexical bundles refer strictly to word sequences with a length of three of more words and are the most frequent sequences in a given register (Biber et al., 1999). *N*-grams can be of any length and are not restricted just to the word level, but can applied to phonemes, syllables, or graphemes as well. This term is are often used in natural language processing (NLP) work to describe such sequences that have been extracted from large corpora. Both lexical bundles and *n*-grams are extracted from corpora, but *n*-grams are a broader type of sequence that encompasses more types of extracted sequences than lexical bundles. *N*-grams are not variable, nor do they have the frequency constraints of lexical bundles. *N*-grams are simply sequences that can be extracted from corpora. Thus all lexical bundles are *n*-grams, but not all *n*-grams are lexical bundles.

This study's focus is on *n*-grams because the broadness of the definition, without strict requirements of frequency, association, semantics, or structure, allows the analyst to select a specific subset of them based on some or all of these constructs. They are very similar to lexical bundles; however, the parameters for selection are more flexible. Moreover, *n*-gram is the terminology used to describe the sequences in the study which I am attempting to replicate. In these foundational studies (Ellis & Simpson-Vlach, 2009; Ellis et al., 2008), the *n*-grams presented to participants were all relatively highly frequent and highly associated. These constitute a special subset of *n*-grams that are widely used by a representative sample of authors across a variety of genres.

Given the subtleties involved in the terminology of the field, I will, in the following section, contextualize several types of FL on four continua to provide a framework for the comparison of different types of FL.

2.1.2 Formulaic Language and Continua of Evaluation

FL, as noted above, comes in so many different forms that examining all the distinctions is beyond the scope of this study (see Wray, 2002, pp. 16-66, for detailed examinations of different taxonomies of FL types). For the purposes of this study I will limit the discussion of FL types to a few particular continua: Fixedness, Semantic Transparency, Frequency, and Association. The first term refers to the degree of formal mutability a sequence has. The second refers to the degree to which the meaning of a sequence is derivable from its constituent words. The final term refers to the frequency of occurrence. Within this framework, I will discuss the characteristics of prayers, song lyrics, idioms, collocations, lexical bundles, and *n*-grams.

The first formulaic language classification continuum is that of Fixedness. Some types of language are particularly fixed. For example, prayers are generally quite fixed and may remain so for hundreds of years under the impetus that the sequence has meta-physical significance, making it a serious cultural violation to alter it. Poems and song lyrics are also generally quite fixed, with some exceptions for live performances or reinterpretation (e.g., Elton John's musical dedication to the late Princess Diana). Collocations, lexical bundles, and *n*-grams are each fixed in the sense that in order to be so classified, the phrase must be re- or occurrent in exactly that form. *N*-grams present an especially interesting case because they are extracted from corpora based on the fact that the sequence occurs. Idioms exist on this continuum as well. Some, such as *by and large* are immutable—changing any part of the utterance disrupts its idiomatic meaning. Other, though, are able to undergo transformations while still retaining their idiomaticity. To borrow an example from Wray (2002), *He spilled the beans* and *The beans were spilled by him* have approximately the same idiomatic meaning. To borrow another such example, *to give [someone] a piece of [someone's mind]* has an idiomatic meaning but allows for a particular set of variant words to be inserted into the slots. However, there are non-idiomatic, frame-slot patterns, such as *Help [pronoun/proper noun] help [pronoun/ proper noun]* (e.g., *help me help her*), that are again only partially fixed. Their variation is often along syntactic lines—i.e., a slot may take any NP or any VP, but not other types of phrases (Wray, 2002).

The second continuum for classifying formulaic language is that of Semantic Transparency. On one end of this continuum is the position of totally semantically transparent (e.g., *the dog chased the skunk*) where the meaning of the sentence given no other information is the sum of its constituent parts. On the other end exist phrases that mean more than the sum of their parts or have a meaning that is only historically related to the words that compose it (as in the case of idioms). Some phrases have both an idiomatic, semantically opaque, and literal meaning (e.g., *Kick the bucket* which means both to die or to literally hit a bucket with one's leg), while others have semi-transparent meaning (e.g., *let the cat out of the bag*), wherein when the cat is out of the bag, the secret being kept is no longer so secret (Wray, 2002). Collocations, lexical bundles , and *n*-grams may fall anywhere on this continuum. For instance, *on the other hand* refers less to literal hands but is used to invoke contrast, while *it has been shown that* is fairly semantically transparent and is used to report on findings in research. An idiom could be a collocation or lexical bundle provided that it sufficiently frequent. A fixed idiom, such as *by and large*, would also by definition be an *n*-gram (contingent on it's attestation in a corpus); however, the more flexible sorts (e.g., *let the cat out of the bag/the cat was let out of the bag*) present a slightly more complicated case. Because these sequences are flexible, each permutation of the sequence would be an *n*-gram in its own right. Moreover, if the sequence is not attested in any corpus, even if theoretically possible, it would not be a n-gram. For instance, using Google N-Gram (Michel et al., 2011), one can search for the phrase *holy Mountain Dew*. This search returns zero results, even though the frame *holy NP* is a recognizable pattern. Although I was able to generate this phrase, it is not attested in the corpus.

Frame-slot patterns may also vary on the continuum of semantic transparency, as some idioms may have this property; however, not all such patterns are idioms. Longer stretches of discourse (i.e., poems, prayers) may vary in relation to the degree that their semantics are readily discernible from the words used to encode the ideas. For example, Carroll's (1872) *Jabberwocky* has considerable semantic opacity, while Key's (1814) *The Star Spangled Banner* may have some degree of semantic opacity or virtually none, depending on the hearer's knowledge of early 19th century US English. Similarly, a nursery rhyme, such as *Little Miss Muffet* may have parts that are both semantically opaque and transparent (e.g., Little Miss Muffet sat on a *tuffet* eating *curds and whey*), while another classic rhyme, *Mary Had a Little Lamb* is far more semantically transparent.

The third continuum of formulaic language classification is that of Frequency. For formulaic language such as a prayer or song lyrics, frequency is not an inherent property of the sequence: These are also often longer, memorized examples of speech that are not typical outside of particular contexts (e.g., religious services, concerts). Another type of FL defined without specific reference to frequency are idioms. Although it might be hypothesized that idioms would be relatively frequent in large corpora, this is not necessarily the case. Conklin & Schmitt (2008) found that even in the 100 million word British National Corpus, idioms were relatively rare. They found that many widely-known idioms did not occur at all or only very infrequently. N-grams have but one frequency requirement: that the sequence in question occurs at least once in a corpus. In contrast, collocations and lexical bundles are defined by their frequency (Biber et al., 1999; Sinclair, 1991).

The final continuum of interest is that of Association. Association in this sense means the degree to which the constituent words of a phrase co-occur beyond the frequency that chance alone would predict. This is to say, a phrase's degree of association is a useful measurement of how strongly the words of a phrase are linked with one another in a corpus. Degree of association can be quantified by several statistics but is predicated on three factors: the frequency of the words and sub-phrases in the collocation, lexical bundle or n-gram, the expected frequency of the phrase given this information, and the observed frequency of the phrase in question (this will be discussed in greater detail below). This statistic is not usually calculated for highly ritualized pieces of language such as prayers, poems, and songs, but is instead usually reserved for sequences extracted from corpora (e.g., collocations, lexical bundles, or n-grams). For instance, to return to the example of *strong tea*, it is a collocation that is not especially frequent, but the words are associated to a degree beyond that which change alone would predict. Similarly, *on the other hand* is one of the most strongly associated n-grams in academic writing. *Than that of the* is associated at a level greater than chance would predict but relative to *on the other hand*, it is only weakly associated. There are two lexical and two functional words in *on the other hand*, while all the words in *than that of the* are functional. Highly associated phrases are often typified by a greater proportion of content words, while more weakly associated phrases are likely to contain more functors. The reason for this is clear: Functors are the most common words in English, thus they are expected to co-occur frequently.

In summary, it is difficult to define formulaicity and FL without describing several continua and evaluating different types of FL on each. Although some types may have their as defining characteristic a position on a single continuum (e.g., collocations and lexical bundles), many types are characterized by variation along several dimensions, there is no single defining criterion in identifying it. Idioms, such as *kick the bucket*, are fixed, semantically opaque, and relatively infrequent.

On the other hand, an idiom that can be considered a lexical bundle or *n*-gram, is fixed, less semantically opaque than kick the bucket, and is quite frequent and highly associated. Another idiom, *X let the cat out of the bag* varies differently. It is not as fixed as *kick the bucket* because it can take variable, animate subjects and be rendered as *the cat was let out of the bag by X* without sacrificing its idiomatic meaning. It is also less semantically opaque than *kick the bucket*, as outlined above. This is not an especially frequent idiom, occurring only 22 times in the 450 million word COCA (Davies, 2008). In contrast to the above idioms, other phrases are variable in different ways. For instance, *strong tea* (as per Durrant & Doherty (2010)) is semantically transparent, formally fixed, and more associated than chance would predict. Not all of these types of formulaic language are easy to identify automatically because computers are not easily to detect semantic transparency (or the lack thereof) or non-adjacent formulae. However, if one narrows the field of inquiry to types of FL that are both immutable and based on frequency, identification becomes much easier, as these properties enable rapid, automated computerized analysis. Collocations, although first coined and described by Firth, became a much more feasible object of study with the advent of computers (Stubbs, 1993). Sinclair's work in the late 1960s on the automated analysis of language made it possible to examine Firth's concept in much greater detail. In the 1990s, Biber et al. (1999) conducted research on lexical bundles and found that 20 to 30% of words in given register are part of these recurrent sequences. In the next section, I will examine the work of these researchers to explicate the basics of corpus linguistics and its link to the production and processing of formulaic language.

2.2 CORPUS LINGUISTICS

The computerized analysis of principled corpora offer a useful for method for identifying formulaic language—e.g., collocations, n-grams, or lexical bundles. Furthermore, corpus-based approaches address a criticism levied at some schools of linguistic scholarship: an over-reliance on contrived or unusual data (Bloor & Bloor, 2004; Sinclair, 1991; Stubbs, 1993). These analytic approaches treat language in use as the object of study, rather than an abstract notion of an underlying competence. The analytic tools of corpus linguistics allow linguists to quantify words, collocations, n-grams, lexical bundles and other lexicogrammatical patterns on a scale that was simply not possible before the digital revolution of the 1980s and 1990s. Computer analysis of large, principled corpora, thus, allows for an arguably more rigorous, empirical way to examine language than either examination of unusual examples or pure human analysis of real text (Sinclair, 1991). The increases in computational power in the first decade of this millennium have made the timely analysis of corpora of hundreds of millions words (e.g., the Corpus of Contemporary American English or the British National Corpus) to billions of words (Google's massive corpus of literature from the past 2 centuries) quite possible, allowing researchers to obtain counts of the occurrences of words, phrases, and grammatical patterns in use.

2.2.1 Pointwise Mutual Information Score

Through corpus analysis, linguists may obtain frequencies of words. The patterns of lexical co-occurrence are what define such structures as collocations and n-grams. However, with the analysis of large corpora, there will be a few extremely frequent n-grams but many that occur less than 5 times Ellis (2012). It follows, then, that corpus linguists need a way to identify which of the sequences are worth further examination. Pure frequency is one way to examine this phenomenon. However, it can also be useful to measure how strongly the constituent words in a phrase are associated with each other. Although a word may be relatively infrequent in a particular corpus, when it does occur, it may be accompanied by another word more often than one would expect given the probabilities of each word (e.g. *strong tea*). One common method of quantifying these associative relations is the pointwise mutual information score (MI) (Church & Hanks, 1990; Pothos & Juola, 2007; Xu, Jones, Li, Wang & Sun, 2007). This measurement is derived from the concept of mutual information in the field of information theory (Church & Hanks, 1990); however, the MI used in corpus linguistics is distinct from information theory's concept of MI (Xu et al., 2007). In this section, I will explain how MI can be calculated for a bigram, discuss issues surrounding the extension of the calculation to multiple words, provide some caveats to its interpretation in corpus studies, and justify the inclusion of MI in my statistical models.

First, it is important to note that MI is not a measure of how certain it is that an n-gram's words are strongly associated; instead, MI serves a measure of magnitude of the strength of association. MI is, thus, an effect size. The MI of a two word n-gram or *bigram* is defined mathematically by the formula:

$$MI_{ab} = \log_2 \left(\frac{O}{pr(a) * pr(b) * W_T} \right) \tag{2.1}$$

where the O represents the observed frequency of the bigram, a and b are the constituent words of the bigram, and W_T is the total number of words in the corpus.

To illustrate this calculation more concretely, I will return Durrant & Doherty's example of *strong tea*. The frequency of this phrase (O) in the British National Corpus was 28 times, and this value is then divided by the expected value ($pr(a) * pr(b) * W_T$), given the probabilities of the words in the phrase (1.26). This calculation results in an MI for *strong tea* that is equivalent to $log_2(28/1.26) = 4.43$.

Church & Hanks suggested that a bigram with an MI of at least 3 is more likely to be a collocation of interest. Since 4.43 > 3 *strong tea* is very likely a collocation.

Second, the calculation of MI can be extended to include sequences of more than two words, as is evident from computer programs such as Collocate (Barlow, 2010), which will calculate MI for sequences of more than two words. However, it is unclear how the program calculates it. As Shaoul & Westbury (2011, p. 185) showed: When it comes to multiword sequences, there are multiple frequencies involved in an n-gram. For instance, in an 4-word n-gram such as *on the other hand*, there are 16 frequencies:

- The frequency of the whole string (1)
- The frequency of the contiguous subgrams (i.e., the contiguous components of the n-gram: *on the, the other, other hand, on the other, the other hand*) (5)
- The frequencies of non-contiguous subgrams (e.g. *on X other hand, on the X hand, on X Y hand, the X hand, on X other*) (6)
- The frequencies of the individual words (4)

The observations of Shaoul & Westbury show that in multiword strings there are many frequencies which could be used to calculate MI. Trnka (2011) and Morgan (2014) suggested that the calculation could be extended to three-word phrases by using the probability of the phrase as a whole, over the product of the individual probabilities. Morgan also proposed that the calculation could be extended to include some of the subgrams but did not include the frequencies of non-contiguous n-grams. I propose that this calculation could be extended to three- and four-word sequences in much the same way but with the inclusion of non-contiguous subgrams. If MI is to be used as a predictor in statistical models of language for multiword sequences, it is essential that researchers know exactly what frequencies are used in the calculation. Do different calculations methods different things? How is it best to calculate measures of association for multiword sequences? Answering these questions is essential so that researchers concerned with this measure of association know exactly what it represents. However, these questions are outside the scope of this thesis. As such, I will use the MI statistic as calculated by Collocate (Barlow, 2010).

Third, in spite of the usefulness of MI as way to quantify the degree of association of the words of a phrase, it is not free of problems. MI scores can become inflated when the frequencies of the phrase low (Church & Hanks, 1990). When the phrasal frequency is low, an alternative way to consider association is through the t-score. This is a test of significance of the association—a t-score of 2 or greater, indicates that the words in question are significantly associated. In contrast to MI, this is not a measure of the strength of the relation between the words. Whereas MI is a measure of magnitude of association, t-scores indicate the level of certainty regarding whether the words in a phrase are in fact associated. The statistic is particularly useful and more reliable when the frequency of the target phrase is less than 5 in any corpus (Church & Hanks, 1990).

Finally, although MI is not a perfect measure of association, I have elected to use it for two primary reasons. First, the selection criteria for the n-grams were intended to eliminate highly discipline or author specific phrases, by requiring that the phrases be in at least 10 texts and 5 different subject areas. Second, as this study was conducted as a partial replication of Ellis and colleagues' work, I elected to use the same measure of association that they had used.

2.3 THE INTERSECTION OF CORPUS LINGUISTICS AND LANGUAGE PROCESSING

With this discussion MI as a marker of collocational and phrasal association and salience in mind, I shall discuss an important development that arose out of Sinclair's work in corpus linguistics. He theorized that language processing and production were governed by two principles: the Principle of Open-Choice and the Principle of Idiom. The Principle of Idiom encapsulates the idea that "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments" (Sinclair, 1991, p. 100). Some language that could be processed or produced under this principle are fixed idioms, collocations, and lexical bundles, which can all have MI and frequency scores associated with them. These statistics would play important part in later finding evidence of whether these phrases are psychologically real (Ellis et al., 2008; Ellis & Simpson-Vlach, 2009; Bannard & Matthews, 2008; Arnon & Snider, 2010). The Principle of Open-Choice, on the other hand, accounts for language that does not appear to be processed like this. This principle accounts for entirely generated phrases, wherein lexical items do appear to be inserted into syntactic frames, as well as phrase like *gimme some X*, wherein *X* represents a slot that can be filled by any noun or NP. Sinclair proposed that most of speech is in fact produced and processed under the Principle of Idiom, with rapid switches to Open-Choice when necessary to fill in slots in frame-slot patterns with individual lexemes. In this way, these principles account for the extensive use of recurrent word combinations, as well as the ability to generate novel language.

For scholars interested in the cognitive aspects of language, corpus analysis provides a new source of data and ways to probe the processes, advantages and limitations of the mind. Do these recurrent patterns represent something in our minds? The answer seems to be, yes (Durrant & Doherty, 2010; Ellis & Simpson-Vlach, 2009; Ellis et al., 2008; Arnon & Snider, 2010; Bannard & Matthews, 2008). In the following section, I will discuss a group of linguistic theories that could account for Sinclair's principles.

2.4 ACCOUNTING FOR SINCLAIR'S PRINCIPLES

In the previous section, I described Sinclair's Principles of Idiom and Open-Choice as a model for language processing and production that arose out of Sinclair's observation of frequent, recurrent phrases in corpora. As I have argued in the previous sections, FL manifests in a range of forms. Some of these sequences are static, ritualized pieces of language that remain unchanged for centuries. They contrast with another extreme: The human capacity to generate entirely novel utterances. One must only consider Lewis Carroll's (1872) *Jabberwocky* to find sentences that are so novel as to be very much open to interpretation beyond their adherence to the morphosyntactic principles of English. However, neither extreme covers all aspects of languages, because some types of language seem to be only partially fixed, e.g., some idioms or sayings (Sinclair, 1991), while other chunks, much shorter than prayers, poems or songs, are totally fixed (e.g., lexical bundles, collocations and some idioms). An important question in linguistics is thus: How can one model of language account for these differences in language production and processing capabilities? A group of theories that fall under a general umbrella of Usage-Based Linguistics could account for Sinclair's principles.

In recent years there has been a substantial interest in FL as a concept at the center of language, rather than the periphery (Wray, 2012). These approaches contrast with analyses of language, wherein a lexicon and a set of combinatorial rules generate language or a set mental sentence templates are filled in by drawing on a lexicon of discrete words with a few fixed phrases that behave like words (e.g., semantically opaque idioms). Examples of such theories of language can be found in Ellis (2002), Langacker (2000), Larsen-Freeman & Cameron (2008), MacWhinney (2000), and Wray (2002). A very general overview of these positions is as follows. First, they rely on prototypes and exemplars that are related in the sense of concept–form–function. These exemplars consist conceptual of non-linguistic information about an entity (e.g., birds) as well as the linguistic forms and functions associated with the forms (e.g. phonology, morphology, names of specific birds, pragmatics of talking about birds, etc.). Exemplars may be conceptualized as *units* which in turn are thought to correspond to sets of neurological structures that activate together in response to a particular stimulus, e.g., seeing a bird or hearing the word *bird* (Ellis, 2002; Langacker, 2000; Rumelhart & McClelland, 1986). It is important to note that the exact patterns of neurons and synapses that constitute these structures that make up units are not known. Units may also be activated when someone recalls a memory of a bird or when someone wants to talk about birds. An important extension of this idea is that exemplars represent the set of neurons that activate in response to any exposure or memory activation of any kind of bird. Among these neural structures are those that correspond to visuo-spatial, auditory, and linguistic memories. In contrast, a rule-based processing system suggests that features such as [+feathered, +beak, +flies, +oviparous] are assembled into the concept of a bird, then linguistic rules are used to combine morphemes into words and insert words into sentences.

Understanding this conceptualization of mental information "storage" is key to the next point: UBL abolishes the difference between conceptualization, phonology, morphology, lexis, and syntax, by considering them part of a singular, larger cognitive apparatus, as opposed to separate stores of concepts, phonemes, lexemes, morphemes, and syntactic frames which are combined through a set of rules. Instead of speech being generated from separate stores of non-linguistic concepts and linguistic parts and combined by rules, UBL posits that when an entrenched neural structure activates, other such structures that are associated with it are preferentially activated. For instance, when someone hears the word *the*, items that succeed *the* (i.e., adjectives and nouns) are the ones primed for greater degrees of activation.

Some UBL theories (e.g., Ellis, 2002; Wray, 2002,0) place FL at the center, while still allowing for a capacity to generate much more novel utterances. The primary driving force of acquisition in these models is the abstraction of linguistic schemata from the input received. This process involves creating a mental map of concept, form, function, and external reality. Whereas some rule-based theories posit that we are in born with syntactic schemata already hardwired into our neurology and apply them deductively, UBL posits that language acquisition is actually an inductive process that abstracts various linguistic and non-linguistic schemata from chunks contextualized input.

Particularly relevant to this study is Ellis' (2002) theory of implicit linguistic tallying. He argued that the frequency of comprehended, contextualized lexical input is driving force of acquisition of words—and by extension, immutable phrases, as well as phonemes, morphemes, etc.—because such input is thought to leave memory traces that increase the sensitivity of the neural pathways associated with the particular word (phoneme, morpheme, lexical bundle, *n*-gram, etc). Thus, under this theory, more frequent words are processed more quickly because the neural pathways are more dredged out. A useful analogy is to consider more frequent words or multiword sequences like highways, while less frequent items are more like unmaintained dirt roads. The highway was built to accommodate a large amount of traffic in the area. It is well-maintained, frequently used, and thus attracts much more use. Conversely an old, hardly used dirt road does not receive much traffic and due to its rough state does not attract much traffic either.

This contrast is particularly important for this paper because these competing theories make different predictions about reaction time (RT) in plausibility judgement tests of n-grams. Some rule-centric theories of language theories predict that there will be no reduction in RT because nearly all language is thought to be generated online, irrespective of frequency or strength of association. On the other hand, UBL theories predict that RT will decrease as a function of frequency or strength of association, depending on the speaker's level of exposure to the target language (i.e., their status as a L1 or L2 speaker).

In the following section, I will discuss studies of language the foundational studies for my work and describe the evidence that Ellis et al. and Ellis & Simpson-Vlach have provided for this position.

2.5 STUDIES OF L1 AND L2 LANGUAGE USE AND PROCESSING

Although Ellis and colleagues have conducted their research relatively recently, their work draws on previous studies (Ellis, 2001,0; Erman & Warren, 2000; Nattinger, 1980; Nattinger & DeCarrico, 1992; Pawley & Syder, 1983; Sinclair, 1991) and is related to other contemporaneous studies (Ädel & Erman, 2012; Chen & Baker, 2010; Conklin & Schmitt, 2008; Durrant & Doherty, 2010; Durrant & Schmitt, 2009; Jiang & Nekrasova, 2007; O'Donnell et al., 2013; Simpson-Vlach & Ellis, 2010). In this section, I shall discuss the literature that examines L1 and L2 speaker use and processing of FL similar to that examined in my study (e.g., collocations, lexical bundles, idioms). Although I will note some of Ellis and colleagues' results, the detailed discussion of those studies will conclude this section and chapter.

In spite of the heavy dominance of rule-based theories of language at the time, Nattinger (1980) recognized the importance of formulaic language in second language instruction. He argued that linguists should give more attention to the role of formulaic language in speech because such language can be more readily tied to particular functions or speech acts. He suggested that ESOL educators ought to base their teaching on the assumption that much of language production is composed of situationally appropriate, prefabricated units. He concluded by proposing that formulaic sequences and their functions “should be the syllabus” for ESOL education (Nattinger, 1980, p. 342).

Following Nattinger, within the fields of both FLA and SLA, there have been two strains of research regarding formulaic language particularly relevant to my study: comparisons of how L1 and L2 speakers use formulaic language and studies of L1 and L2 speakers’ processing of formulaic language.

2.5.1 Use of FL

Corpus-based analysis in particular has lent support to Nattinger’s argument by making it clear that there are frequent, recurrent, and seemingly fixed patterns in speech and writing. Biber et al. (1999) found that between 20 and 30% of words in L1 speakers’ linguistic production across registers are in these lexical bundles. Using the methods developed by Biber and his colleagues, some researchers have examined how second and foreign language learners of English use lexical bundles. The few studies that compare and contrast L1 and L2 speaker use of *n*-grams—and collocations, lexical bundles, etc.—by means of corpus analysis have focused on use in English academic writing (Ädel & Erman, 2012; Chen & Baker, 2010; Durrant & Schmitt, 2009; Siyanova & Schmitt, 2008). Chen & Baker (2010) and Ädel & Erman (2012) compared the usage of 4-word lexical bundles in the academic writing of L1 and L2 speakers. Each of these studies found that although both L1 and L2 speakers used lexical bundles, a relatively small proportion of these bundles are shared by the two groups. Moreover, both groups of researchers showed that L1 speakers used a greater variety of lexical bundles than L2 speakers. These studies showed that L1 and L2 speakers both use lexical bundles in their writing. However, neither sought to explain from a cognitive perspective why so few are shared by these groups. In contrast, Durrant & Schmitt (2009) conducted a similar study but did suggest some implications for cognitive development in SLA. The researchers identified 2-word collocations in order to investigate their use in the writing of L1 and L2 speakers and, like Ädel & Erman and Chen & Baker, found that L1 speakers used a greater variety of the target FL than L2 speakers. The other most significant findings of this study were that the L1 speakers used more low-frequency collocations than L2 speakers and that the L2 group tended to underuse collocations with a low frequency but high MI score. Concluding that second language learners effectively acquire high-frequency collocations in the target language, the researchers also noted that learners take longer to acquire less frequent but strongly associated items. Durrant & Schmitt's findings support Ellis and colleagues position that L2 speakers more readily acquire high-frequency formulaic sequences rather than low-frequency sequences that have a high degree of association

However, not all studies have found this seeming disparity in the use of collocation or extended collocational structures by L1 and L2 speakers.For example, Siyanova & Schmitt (2008) found that in English-language essays written by Russian EFL learners, approximately 45% (p. 435) of learner collocations were attested in the British National Corpus (BNC), while a approximately 25% did not occur, and another 25% occurred only infrequently. Non-significant differences were found in the use of collocations in similar essays written by L1 speakers of English. The researchers concluded that the writers in their sample used approximately same amount of collocations and note that this “does not support a commonly held view that L2 learners underuse native-like collocations” (Siyanova & Schmitt, 2008, p. 439). A finding that came out of Siyanova & Schmitt's study was that although L2 speakers used approximately the same number of “native-like” collocations, they were not as accurate as L1 speakers in judging how frequent a given collocation is. The researchers showed that while use between the groups may be comparable, the processing mechanisms underlying use are different for the two categories of speakers.

These studies show that comparisons of FL use by L1 and L2 speakers of English only tells part of the story of how these sequences may be represented in the mind. In order to better understand the less visible processes of SLA, it is necessary to examine some studies which have considered the processing of FL.

2.5.2 Processing of FL

As is evident from the previous section, there has been some interest in formulaic language use by both L1 speakers and L2 speakers. However, there has also been interest its processing by these same groups. Some researchers have examined only how L1 speakers process various types of phrasal FL language (Arnon & Snider, 2010; Bannard & Matthews, 2008; Durrant & Doherty, 2010; Ellis & Simpson-Vlach, 2009), while other researchers, using L1 speakers as a comparison group have considered how L2 speakers process this sort of language (Conklin & Schmitt, 2008; Durrant & Schmitt, 2009; Ellis et al., 2008; Jiang & Nekrasova, 2007; Siyanova & Schmitt, 2008). A few studies have examined whether categorical definitions of formulaicity are associated with reduced processing time, while other studies have consider whether the frequency of a phrase is associated with reduced processing time. Very few, though, have examined whether the degree of association of a phrase influences processing time.

In the following section, I will discuss several studies that have examined formulaic language processing by L1 and L2 speakers of English. Studies in FLA and SLA have found that high frequency phrases are generally processed more quickly than low frequency phrases. Similarly, there is evidence that for L1 speakers of English, MI rather than frequency may be the more appropriate predictor of increased processing speed (Ellis & Simpson-Vlach, 2009). In the following paragraphs, I will review studies in these two veins of research.

FLA

One group of studies from the FLA literature has shown that highly frequency multiword sequences may be processed more like complex, extended lexemes rather than generated phrases. These studies suggest that processing advantages for high-frequency sequence is present in both child and adult L1 users of English. For instance, using paired stimuli derived from a corpus of child-directed speech, Bannard & Matthews (2008) found that L1 English-speaking children were more likely to successfully repeat a frequent four-word sequence than a paired less frequent one. (Arnon & Snider, 2010) employed a similar stimulus generation strategy and found significant frequency effects for L1 adult processing of these 4-word sequences: L1 speakers tended to judge highly frequent sequences more quickly than less frequent ones. Durrant & Doherty (2010) considered the effects frequency and psychological association on the linguistic processing of L1 speakers an attempt to separate the effects of priming arising out of high collocational frequency and semantic association. They concluded that their results support the idea that high-frequency collocates are to some degree psycholinguistically valid but noted (p. 145) that collocational priming effects may only exist for phrases which show a high measure of statistical association. Ellis & Simpson-Vlach (2009) identified 3-5 word *n*-grams in academic writing and found that MI, not frequency, was the primary predictor of reduced reaction in L1 speakers of English. These studies suggest that L1 speakers have some sensitivity to various measures of formulaicity, such as threshold frequency measures (e.g., Arnon & Snider (2010); Bannard & Matthews (2008); Durrant & Doherty (2010)), and phrasal association (Ellis & Simpson-Vlach, 2009). In this next section, I will examine studies conducted to better understand the processing of formulaic language in the context of SLA.

SLA

Within SLA, experimental studies have provided evidence for the psychological reality of categorical measures of formulaicity (e.g. idiomaticity) (Jiang & Nekrasova, 2007; Conklin & Schmitt, 2008) and quantitative measures of formulaicity such as frequency and degree of association (e.g., Durrant & Doherty, 2010; Ellis & Simpson-Vlach, 2009; Ellis et al., 2008; Siyanova & Schmitt, 2008). Another group of studies from SLA literature has shown that a sequences degree of formulaicity may influence how quickly L1 and L2 speaker of English process multiword sequences. Employing corpus-derived lexical bundles identified by Biber et al. (1999) and grammatical but non-formulaic counterparts (e.g., *to tell the truth* versus *to tell the price*), Jiang & Nekrasova (2007) conducted two experiments in which they found that L1 and L2 speakers processed the formulaic test items significantly more quickly than non-formulaic sequences. In a similar study using high-frequency idioms, Conklin & Schmitt (2008) found that both L1 and L2 speakers processed idioms significantly more quickly than non-idiomatic, non-formulaic counterpart items. These studies suggest that certain categorical definitions of formulaicity, such as idiomaticity or formulaicity in the form of lexical bundles that appear salient to the researcher, impart some processing advantage for L1 and L2 speakers of English.

Another group of studies have employed more specific quantitative measures of formulaicity to examine the effects of frequency on RT. For instance, Siyanova & Schmitt (2008) showed that both L1 and highly-proficient L2 English speakers recognize more frequent adjective-noun collocations than less frequent ones. However, they concluded based on the three studies conducted that while L2 speakers may use a comparable amount of frequent collocations to L1, they have less fine-grained intuition about the frequency of the collocations under examination. Ellis et al. examined the role of frequency and degree of phrasal association on the processing of 3-5 n-grams by L1 and L2 speakers. They found that MI predicted reduced reaction RT in L1 speakers, while frequency was the key predictor of reduced RT for L2 speakers. These studies suggest that L1 and L2 speakers may have different processing mechanisms for highly associated n-grams.

Although the previous studies have focused on both use and processing of formulaic language in a second language context, a gap exists in this research: namely, only Ellis and colleagues have also consider the degree of association directly as a predictor of reduced RT in addition to frequency. While some studies in these fields have made use of MI in stimulus generation process, most have not explicitly examined if this measurement is indicative of a processing advantage. For example, Durrant & Doherty (2010) stratified their stimuli by frequency, MI, and t-scores bands, but these measurements were not included in their statistical models. Durrant & Schmitt (2009) inferred some possible frequency effects and mentioned the tendency of learners to underuse sequences with high MI scores; however, the initial focus of their research was not to look for possible frequency or association effects. Siyanova & Schmitt used MI as a part of their stimuli generation criteria but did not consider its effects on processing time. Clearly, MI has mostly been exploited as a selection criterion, rather than as a predictor of reduced processing latency. Moreover, as O'Donnell et al.'s (2013, p. 86) literature review shows, Ellis and colleague's findings on polylexical n-grams have not been replicated by other researchers. Thus, my study represents an attempt to further examine the phenomenon identified by Ellis et al.. In order to explicate some of the rationales behind my study's methodology, I will review in more detail the work Ellis and colleagues in the following section.

2.6 FOUNDATIONAL STUDIES FOR REPLICATION

Ellis, Simpson-Vlach & Maynard (2008) and Ellis & Simpson-Vlach (2009) attempted to unify the areas of formulaic language within the context of SLA, corpus linguistics, and psycholinguistics, by examining L1 speakers' and L2 speakers' processing of n-grams of varying lengths, associational strengths, and frequencies. Their results provided evidence that strength of association is the factor that most affects reduced processing time for L1 speakers, while frequency is the predictor that most affects a reduction in reaction time for L2 speakers.

The purposes of these studies were quite similar. Both studies were concerned with establishing the psycholinguistic reality of formulaic language that was found through the analysis of corpus-based statistical measurements—in particular frequency and MI. Each set of researchers examined what correlational relations, if any, these measurements might have with language processing. Ellis et al. included both highly proficient L1 and L2 speakers who were enrolled in graduate degree programs whereas Ellis & Simpson-Vlach (2009) were concerned with establishing this reality only for L1 speakers of English. The methodologies of these studies also shared many of the same features but diverged in key ways. I will discuss three major similarities in these studies' designs. First, the researchers identified n-grams of 3- to 5-words (e.g., *on the other hand*or *the extent to which*) in academic writing and evaluated them for both frequency and MI value. They then grouped them into nine categories based on frequency and MI (e.g., low-frequency/low-MI, medium-frequency/low-MI, high-frequency/low-MI). Second, they asked experienced ESL teachers to rate these sequences on a Likert scale for three different measurements: perception of formulaicity, perception of semantic or functional cohesion, and perception of pedagogical value. Each measurement of perception had a significant positive correlation with the others (e.g., a high perception of educational value and pedagogical usefulness correlated with perception of formulaicity). Using further correlational analysis, the researchers found that although frequency and MI contributed to the instructors' valuation of the n-grams, MI most influenced their prioritization.

Finally, the researchers in each study employed three of the same quasi-experimental techniques. For each experiment, the researchers used 108 n-grams as test items.

The independent variables were n-gram length, frequency, MI, and in Experiments 2 and 3, the phonemes per n-gram. Each item was associated with different values for these variables. Examples of these test items include *the extent to which* (highly associated, medium frequency), *the value of the* (medium association, low frequency), and *in the context of the* (low association, low frequency). Ellis et al. (2008) operationalized processing time by response time to the stimuli (either time to press a button or voice onset time after the appearance of the item. They subsequently used forced-entry multiple regression to find any statistical effects of n-gram length, frequency, and MI on RT. Ellis & Simpson-Vlach and Ellis et al. found that increased MI was the significant predictor of reduced RT for L1 speakers, while increased frequency was the significant predictor of reduced RT in L2 speakers. Increased n-gram length was associated with significantly increased RT in both L1 and L2 speakers.

In the places where these studies share methodologies, the researchers found similar results. In each study, the researchers found that MI was the strongest predictor of reduced processing time for L1 speakers; however, frequency was never found to have a significant negative correlation with these measurements. In contrast, Ellis et al. found that n-gram frequency was the key predictor of RT reduction for L2 speakers.

Ellis and colleagues interpreted these results as being consistent with Ellis' theory of implicit linguistic tallying (cf. p. 23, this document). Both groups of researchers noted that L1 speakers were sensitive to MI and that this phenomenon must indicate some unconscious representation of frequency because MI is in part derived from the overall frequency of each word in an n-gram. In order to be sensitive to this property of n-grams, L1 speakers must have an overall sense lexical frequency. Contrastively, they concluded that since L2 speakers had not had nearly so much exposure to the language, they had only developed sensitivity to units that were frequent but not to low-frequency bundles that have a strong association between the words.

Further inquiry into this area could show interesting results related to the processing of *n*-gram by L1 and L2 speakers. The inclusion of L2 speakers will provide another data point regarding the role of frequency in second language acquisition. Statistically significant results would further the case for the role of frequency in language acquisition and support Ellis's (2002) theory of implicit linguistic tallying. Such results could also provide important evidence that could inform how language educators construct syllabi. In the following chapters, I will examine this phenomenon by considering the following research questions:

- Do the predictors of L1 versus L2 speaker status, *n*-gram frequency, and *n*-gram MI significantly influence reaction time (RT) on plausibility judgment tests of highly frequent, highly associated *n*-grams?

Ellis et al.'s (2008) and Ellis & Simpson-Vlach's (2009) work has several implications that have informed my research hypotheses, enumerated here.

- L1 speakers will process high-MI *n*-grams more quickly than low-MI *n*-grams, irrespective of their frequency.
- L2 speakers will process high frequency *n*-grams more quickly than low frequency ones, irrespective of the degree of association.

Chapter 3

METHODS

In this chapter, I will discuss the factors necessary to test my hypotheses and answer my research question including: the participants, the corpus from which the stimuli were drawn, the software used to extract the test items, the demographic survey that participants completed, the measurements of interest and covariates, the experimental procedure, as well as a justification for departing from the statistical methods employed by Ellis, Simpson-Vlach & Maynard (2008).

3.1 RESEARCH DESIGN

This experimental study examined the Reaction Times (RTs) of L1 and L2 English speakers' judgments of one type of formulaic language (FL)—frequent, highly associated *n*-grams. Since participants made repeated judgments, the data are necessarily nested and multi-level. Independent variables associated with each stimulus item were included in the statistical model at level 1. These phrases consisted of three levels of frequency (Low, Medium, and High) and three levels of MI (Low, Medium, and High). Independent variables associated with each participant are entered into the statistical model at level 2. Each participant was associated with a single categorical predictor of RT: L1 or L2 speakerhood.

3.2 PARTICIPANTS

Two groups were the primary focus of this experiment: high proficiency L1 and L2 speakers of English. Selection of participants included recruitment in classrooms, personal contacts, and snowball sampling. Each group consisted of speakers with at least some graduate education in the physical or social sciences. It was necessary to draw participants from these disciplines because the corpus from which the stimuli were drawn contains texts from those disciplines, and speakers in other disciplines may not have encountered this type of language. The minimum goal for sample recruiting was set at 20. Kreft & de Leeuw (1998) suggested this value to be the minimum sample size needed to detect in cross-level interactions in a Mixed Effect Model (MEM). In their discussion of power, it is unclear whether 20 sample is the minimum to detect a medium or large effect for the cross-level interaction. As Kreft & de Leeuw noted, the power of a model is is largely influenced by the strength of the effect.

There were 13 L1 speakers of English who participated in this study. Participants in L1 speaker group range in age from 25 to 72, with levels of graduate education or a combination graduate education and professional work in academia ranging from 6 month to 50 years. These speakers had lived in primarily English speaking environments for between 25 and 72 years. In this sample were 11 women, 2 men, and no participants identifying as trans or a non-binary gender.

The second group consisted of 10 late bilingual speakers of English with graduate education ranging from 1 month to 14 years. In this sample were 7 women, 3 men, and no participants identifying as trans or a non-binary gender. A variety of first languages were represented in the L2 speaker group including Arabic, Dari, Mandarin, Korean, Polish, Russian, Berber (Tamahaq), Taiwanese, and Thai. L2 speakers had lived in primarily English speaking areas in time-spans ranging from 1 month to 18 years. These speakers must have acquired English when they were past the age of 13 because late bilinguals show different patterns of neurological activity than early bilinguals in grammaticality judgment tests when fluency is controlled for (Abutalebi, Cappa & Perani, 2001). Participants demonstrated high proficiency in English by meeting the university requirements on the TOEFL (or equivalent English language proficiency test) or by being in a graduate level ESOL glass. Three participants reported internet-based TOEFL (iBT) between 83 and 105. Five participant reported paper and electronic TOEFL scores between 530 and 580 (Portland State University, 2014). Some L2 participants scored below the 550 university cutoff, but they were all from a graduate-level intensive English language course or had already earned a graduate degree. One participant reported a 6.5 on the IELTS and one participant did not have to take the TOEFL to obtain admittance to PSU because of her residency status.

Table 3.1: Descriptive Statistics by Speaker Group

	L1			L2		
	Mean	SD	Range	Mean	SD	Range
Age (years)	38.93	14.3	25 - 72	32.78	4.47	23 - 38
EN Environment (years)	36.04	15.2	25 - 72	5.03	4.87	0.1 - 14
Academic Experience (years)	11.39	16.05	0.5 - 50	5.58	4.61	0.33 - 14
Academic Reading (Hrs/wk)	8.71	10.78	0 - 35	17.7	13.02	0 - 40

Selecting high-proficiency English speakers of both groups was done to ensure that the participants would have completed a substantial amount of academic reading in English. They would, thus, be more likely to have encountered the *n*-grams under investigation in their own reading and writing, increasing the probability that they would comprehend the *n*-grams as well as be sensitive to the effects of frequency and/or MI. Such exposure may have allowed more opportunities to see these *n*-grams in context and to make mental form-meaning mappings that would facilitate sensitivity to frequency or MI. Since the study is based on the idea that these corpus-derived statistics (i.e. frequency and MI) may be a factor in reduced RT, it is important that participants have had enough contextualized exposure to the type of writing used for formulating the test items in order for these effects to have developed. Moreover, it was difficult to control strictly for factors such as amount of English exposure, length of study, academic field, and first language (L1) because of the limited pool of possible participants.

3.3 ITEM SELECTION

These n-grams were drawn from the Longman Corpus of Academic Writing. This 5.4 million word corpus contains 75 different texts from 13 fields in the physical and social sciences as well as humanities (Biber, Johannson, Leech, Conrad & Finegan, 1999, pp. 32-33). The majority of these texts were peer-reviewed journal articles written for technical audiences; however, some texts from textbooks were included. I chose to use this corpus for 3 reasons. First, it is a principled corpus, containing a representative sample of academic writing in the physical and social sciences. Thus, the corpus is less likely to contain lexico-phrasal idiosyncrasies particular to one field. Second, the corpus is at least 15 years old, making it less likely that most participants would have actually encountered any of the texts. Thus, it is less likely that any processing effects observed would have been due to participants—who had largely entered academia in the last decade—having actually read the texts in the corpus. Finally, I had access to the corpus in the form of text files that can be analyzed with Collocate (Barlow, 2010). This is in contrast to a data set such as the Corpus of Contemporary American English (COCA) (Davies, 2008). The COCA is primarily accessible through a less flexible web interface. Although it is, at the time of writing, possible to purchase access to the COCA, at the beginning of this study, it was not possible.

N-grams were identified from the Longman Corpus of Academic English with Collocate (Barlow, 2010), a program that allows the user to load textual corpora and set parameters for the extraction of n-grams. The program returns a list phrases with the specified length, along with absolute frequencies, and mutual information scores (MI).

Table 3.2: Descriptive Statistics of n-grams

	Mean	SD	Range
MI	12.7	5	4.23 - 21.78
Frequency (per million words)	18.53	13.98	10 - 116.04

After identifying these bundles with Collocate, the n-grams were aggregated and sorted, so a random sample could be selected. Items needed to occur in a minimum of different 10 texts in the corpus, so as to exclude items that are either highly person- or discipline-specific. The minimum frequency of any test item was 10 instances per million words (Ellis & Simpson-Vlach, 2009; Ellis, Simpson-Vlach & Maynard, 2008). Under these parameters, approximately 350 4-word n-grams were identified. In contrast to Ellis et al. (2008) and Ellis & Simpson-Vlach (2009), n-grams of 3- and 5-word lengths were not extracted, as those studies showed consistent effects for word length. In each of the 7 experiments detailed in those two studies, longer n-grams were associated with significantly slower reaction times. Given the consistency of these findings, I determined that length was not a variable that needed to be tested and so used only only 4-word n-grams as stimuli.

I then divided the stimuli into 9 categories based on high, medium, and low MI as well as high, medium, and low frequency to mirror Ellis et al.'s stimulus categories. I determined the ranges of the categories from the visual analysis of histograms of the frequency and MI of all the extracted n-grams and the mathematical analysis of their distributions. Table 3.3 shows the ranges of values in each category. The MI values were suitably normal that I used 1 SD from the mean as the upper and lower bounds of the Medium category. Values below or above these cutoffs were considered Low or High MI respectively. As expected, the frequency values were heavily positively skewed, so I calculated the semi-interquartile range on the data. This resulted in the upper and lower bounds used to divide the potential stimuli.

Table 3.3: Prespecified Ranges of Frequency and MI

	Low	Medium	High
Frequency (per million words)	10 – 11	11.5 – 18.7	19.6 – 116
MI	4.23 – 7.49	7.5 – 16.49	16.5 – 21.04

For each of the nine frequency/MI distinctions, 9 *n*-grams were identified for a total of 81 such test items, so as to mirror Ellis et al's (2008) design. However, two categories—LL and HL—did not have at least 9 examples from the sample of eligible *n*-grams. In order to balance the design, I identified the three lowest MI *n*-grams in ML category and put them into the LL category. I then selected the 7 highest MI phrases from this category and placed them into the HL category for the balance. Thus, the maximum MI in the LL was 8 and the minimum MI in the HL group was 17. Examples of the stimuli can be found in Table 3.4.

Table 3.4: Examples of Experimental Stimuli

	Low MI	Med MI	High MI
Low Frequency	*than that of the*	*is related to the*	*the structure of the*
Med Frequency	*the effect of the*	*as a matter of*	*a great deal of*
High Frequency	*has been shown that*	*the extent to which*	*on the other hand*

In order to create distractor items, I wrote a simple R script to scramble the stimuli into 81 ungrammatical lexical sequences (R Core Team, 2014), so that participants would not simply rate all sequences as plausible. When the output of the script resulted in a grammatical phrase, it was applied again so that no distractor items would be grammatical. The stimuli fall into binary categories: attested phrases—and thus plausible—and ungrammatical phrases, which are much less likely to be judged as plausible. Examples of these distractors can be found in Table 3.5

Table 3.5: Examples of Experimental Distractors

	Low MI	Med MI	High MI
Low Frequency	*of the that than*	*to the is related*	*of the structure the*
Med Frequency	*the of effect the*	*a of matter as*	*great deal a of*
High Frequency	*been has that shown*	*which the to extent*	*other the on hand*

3.4 DATA COLLECTION

Participant RTs were collected with PsychoPy (Peirce, 2009), participants were be shown the test items in a random but pre-determined order. Each participant was shown a different random order so that no extra dependency was introduced into the model. Additionally, in an attempt to control for possible effects of hand-dominance, the indicator buttons for 'plausible' and 'implausible' were rotated between participants.

At the beginning of the experiment, participants were shown instructions for the task: "Indicate as quickly and as accurately as possible whether the sequence of words you see is plausible in English. For instance, a phrase like 'on the phone' is more likely to occur than 'the on phone'. Use [key] to indicate that the phrase is likely to be seen in English. Use [key] to indicate that it is not." The participants were asked to make plausibility judgments as quickly and accurately as possible about the *n*-grams, by pressing either the *z* or the *m* buttons to indicate whether or not the items were plausible.

Stimuli were presented in white sans-serif font on a grey background. There was an automatic 500 ms delay between stimuli, as pilot testing indicated that requiring the participant to explicitly move to the next stimulus item was disruptive to timely completion of the task. During this delay, the screen was blank before the stimulus appeared. Having read the directions, participants were shown a series of ten practice items, before being presented with the directions a second time. Participants were allowed to take as much time as they desired between the completion of the practice items and the commencement of the official judgement task. PsychoPy recorded each participant's RT for the items. The data then were cleaned in SQLite3 in preparation for analysis in R (R Core Team, 2014). The cleaning process entailed the coercion of data to the appropriate R Type (e.g., integers, real numbers, etc.), the removal of outliers, incorrect judgments, and judgments made about distractor items. Packages used for cleaning were *sqldf* (Grothendieck, 2014), *doBy* (Højsgaard, Halekoh, Robison-Cox, Wright & Leidi, 2013), and *gdata* (Warnes, Bolker, Gorjanc, Grothendieck, Korosec, Lumley, MacQueen, Magnusson, Rogers & others, 2014). Additionally, I calculated the log transformation during the cleaning process. Using the lme4 package (Bates et al., 2014), the data were then fit with several mixed-effects models to find the model of most theoretical and statistical appeal. All graphs were constructed with the ggplot2 package (Wickham, 2009).

3.5 MEASURES AND COVARIATES

In this section, I will operationalize the measures and covariates of this study. I shall first describe the dependent variable, the natural logarithm of Reaction Time (RT), before considering the independent—predictor—variables. For the purposes of this study, the independent variables were divided into two categories: Level-1 and Level-2 predictors of RT. Level-1 predictors (frequency, level of association, and rank in the order of presentation) are measurements associated with each stimulus item, while Level-2 predictors (L1 or L2 speaker group and Personal Mean Reaction Time) are characteristics of the participants.

Table 3.6: Comparison of Variables

	This study	Ellis et al.	Notes
Level-1 Variables			
MI	Interval	Interval	–
Frequency	Interval	Interval	–
Length	–	Interval	Present only Ellis et al. (2008)
Order in Presentation	Interval	–	Covariate in this study
Level-2 Variables			
Speaker Group	Nominal	Nominal	–
Personal Mean RT	Ratio	–	Covariate in this study
Dependent Variable			
RT	Ratio	Ratio	Log transformed in this study

3.5.1 Dependent Variable

The dependent variable was the natural logarithm of RT. Only items appropriately judged as plausible were included in the statistical analysis. Approximately 14% of responses were lost due to the judgement of an ungrammatical phrase as plausible. RTs less than 200 ms and greater than 3 SDs above a participant's mean were excluded as outliers. RT was measured in milliseconds (ms), making it a continuous, ratio measurement. It was, however, necessary to log transform the data because participants could take as long as they needed to make a judgement, which led to a marked positive skew to the data. In the models I tested, using raw scores in milliseconds results in a non-normal distribution of residuals, which violates the assumption of residual independence and homoscedasticity. By regressing the log transformed DV on the predictors, the residuals were much closer to the Gaussian Normal distribution.

3.5.2 Level-1 Predictors Associated with each Stimulus Item

In this section, I will describe the measurements and covariate associated with each stimulus item. The first measurement of interest is the *n*-gram's Mutual Information score (MI). MI is a continuous measurement that quantifies the magnitude of association between words in a phrase Church & Hanks (1990), with greater values indicating stronger association. The *n*-grams in my sample range in MI from 3 to 22. Ellis et al. (2008) and Ellis & Simpson-Vlach (2009) found this statistic to be a significant predictor of reduced reaction time in L1 speakers of English. Second, frequency is a continuous measurement of a stimulus item's occurrence within the corpus. The values of this variable in my data ranges from 53 to 615. Ellis et al. (2008) and Ellis & Simpson-Vlach (2009) found this to be a significant predictor of reduced reaction time in L2 speakers of English. Finally, in order to control for possible task learning effects, the rank of the stimulus item in the order of presentation was considered for each participant. This covariate is a continuous measurement that ranged from 1 to 162.

3.5.3 Level-2 Predictors Associated with Individuals

The primary Level-2 predictor is L1 or L2 speaker group. Examining this predictor in a cross-level interaction is intended to examine whether the speaker groups exhibit different patterns with respect to the Level-1 predictors of MI and frequency. The continuous measure of Personal Mean Reaction Time (PMRT) serves as a continuous, non-occasion varying Level-2 covariate. Although this variable is similar to speaker group, they are not the same. As Table 3.7 (p. 46) shows, the range of RTs for each group overlaps, suggesting that beginning to acquire a language at birth is to some degree correlated with faster average processing; however, this status is not the only factor determining a participant's PMRT. There may be other processes at play, such as age, motor coordination, visual acuity, or visual processing speed that might account for this overlap. Including this covariate in the statistical models mathematically controls for some of the other unmeasured factors that may contribute to faster or slower processing.

Table 3.7: Group Summary of Reaction Times (RT)

	Minimum	Maximum	Mean	SD
L1	0.90	1.93	1.351	0.278
L2	1.58	3.27	2.201	0.438

3.6 STATISTICAL METHODS: MIXED-EFFECTS MODELING

In this section, I will introduce Mixed-Effects Models (MEMs) and review the advantages of choosing a MEM for these data, as opposed to the Aggregate Ordinary Least Squares analysis that Ellis and colleagues employed to analyze their data.

3.6.1 Mixed-Effects Models: An Overview

Mixed-effect modeling is a multivariate regression technique that is especially useful for the analysis of data consisting of repeated measures. This technique enabled me to fit a single model to the data that allowed for the examination of each participant's individual patterns of response to different types of stimuli, in addition to patterns present across the L1 and L2 speaker groups. This analytic method contrasts with the Aggregate Ordinary Least Squares (AOLS) analysis that Ellis and colleagues used. The AOLS analysis involved averaging reaction times for each item across participants. The results of an AOLS model, on the other hand, only provide information about patterns across groups.

3.6.2 Mixed-Effects Models: Advantages

Although it is mathematically possible to analyze these data with ordinary least squares regression (OLS) techniques, e.g., Disagreggate OLS, Aggregate OLS, and Disaggregate OLS with Dummy Coding, a MEM is a more appropriate choice of analysis (Cohen et al., 2003). First, the data consist of repeated measurements of RT from individuals, which violates the assumption of independence of observations in OLS. If the analyst does not take this sort of clustering into account, the estimates of coefficient standard errors are negatively biased, resulting in α-inflation (Cohen et al., 2003; Kreft & de Leeuw, 1998; Singer & Willet, 2003). The Intraclass Correlation Coefficient (ICC) was calculated to be 0.276 (Revelle, 2014) and indicated non-trivial clustering in the data (Cohen et al., 2003; Kreft & de Leeuw, 1998). Thus, there is quantitative justification for proceeding with the MEM.

Second, MEMs simplify the analysis of data with missing values because they do not require a completely balanced design. For this study, it was expected that participants would not be 100% correct in their judgments; moreover, it was a virtual certainty that participants would make incorrect judgments on different items. With a MEM, it is possible to exclude incorrect judgments that occur on different items without invalidating the model.

Third, MEMs treat dependency as a feature of the data, rather than a problem to solve. An Aggregate OLS analysis—the method that Ellis et al. employed—attempts to correct for this by averaging the variables over participants within groups; however, this approach comes a price. In this study, this price would be the sacrifice of the individual distributions of RTs and the implicit assumption of a uniform base reaction time and predictor-effects across all participants. This approach is akin to seeing the forest but being unable to see any particular tree in much detail. As Robinson (1950) demonstrated, sometimes correlations at lower levels of aggregation may be radically different from the trend over all groups.

Moreover, This method may also positively bias the coefficients' standard errors, leading to an increased Type II errors (Cohen et al., 2003). Employing this analytic method could lead to the failure to reject H_0 when it is, in fact, false.

A MEM, in contrast, allows for the examination of intra- and interpersonal variability in RT. In my study, this technique will allow me to see how particular individuals reacted to the stimuli in addition to how the L1 and L2 speaker groups reacted as wholes, making interpretation of the parameter estimates much clearer: With a MEM, it is neither necessary to sacrifice the individual distributions in an aggregation nor use dummy codes to account for interpersonal variation in RT. Employing a MEM allows the analyst to ascribe variation in the individual parameter estimates to differences in another predictor.

Finally, a single MEM can be fit to the data to avoid having to fit a multivariable regression model for each group of speakers. In Ellis et al.'s analysis, an OLS regression analysis was conducted by regressing RT on the predictors or MI, Frequency, and words per phrase. Speaker status was not included in either linear model as a predictor, as a separate OLS regression was conducted for the aggregate scores of each group. Although, the researchers did find significant results for MI in the L1 English speaking group and for frequency in L2 speakers, this approach leads to a near doubling of the Type I error rate, since each analysis has a Type I error rate of 5%. Using a single model controls for this sort of α-inflation.

Chapter 4

RESULTS

In this chapter, I shall discuss the results of my study. The results suggest that highly associated *n*-grams are psychologically real for L1 speakers. They also suggest that highly frequent *n*-grams are psychologically real for both L1 and L2 speakers.

4.1 MODEL SPECIFICATIONS AND RESULTS

I will describe the model selection process, specify my model mathematically, and then present the results of my final model, Model D. The analytic process entailed testing a variety of models in order to see which best fit the data (Bates, Maechler, Bolker & Walker, 2014). This process required modifying the fixed effect structures of the models, thus requiring the MEM to be fit through Maximum Likelihood (ML) estimation (Kreft & de Leeuw, 1998; Singer & Willet, 2003). Fitting the models with Restricted Maximum Likelihood (REML) would have been inappropriate because the fixed effects structure was changed for some of the fitted models, thus rendering them incomparable.

In this section, I will provide information on interpreting the coefficients and report on the findings of each model. I have chosen not to include specific numeric values in the text itself but to display them in tabular format (cf. Table 4.1, p. 50) because the actual values of the regression coefficients are not especially meaningful on their own. Since there is no direct conversion of the coefficients milliseconds, I will wait until the next chapter to discuss the practical numerical results that arise

Table 4.1: Table of Results

	Effects	Parameter	MEM A	MEM B	MEM C	MEM D
π_{0i}	Intercept	γ_{00}	7.3593***	7.4944***	7.549***	7.5477***
			0.0577	*0.0603*	*0.0354*	*0.0354*
	L1	γ_{01}	–	-0.3565*	–	–
			–	*0.0802*	–	–
	PersMeanRT	γ_{02}	–	–	0.3999***	0.3972***
			–	–	*0.0369*	*2e-04*
π_{1i}	MI	γ_{10}	-0.0012	0.0044	0.0051	0.0048
			0.0026	*0.0027*	*0.0033*	*0.0031*
	L1	γ_{11}	–	-0.0102*	-0.0112*	-0.0059*
			–	*0.0036*	*0.0039*	*0.0028*
π_{2i}	Frequency	γ_{30}	-5e-04*	-5e-04*	-5e-04*	-5e-04*
			1e-04	*2e-04*	*2e-04*	*1e-04*
	L1	γ_{31}		1e-04	1e-04	–
			–	*2e-04*	*2e-04*	–
π_{3i}	Order	γ_{40}	-8e-04*	–	-8e-04*	-5e-04*
			2e-04	–	2e-04	2e-04*
Variance Components						
Level-1		σ^2_ϵ	0.3465	0.3507	0.3463	0.34657
Level-2		σ^2_{int}	0.0651	0.0298	0.0104	0.0103
		σ^2_{MI}	8.7e-05	4e-06	6.7e-05	6.3e-05
		σ^2_{Freq}	1e-08	2.3e-10	2e-08	–
Goodness-of-Fit						
AIC			1366.5571	1370.2782	1312.8148	1306.3281
BIC			1426.5399	1441.167	1389.1565	1360.8579
Deviance			1344.5571	1344.2782	1284.8148	1286.3281

from Model D, the final, best-fitting model.

4.1.1 Notes on Interpretation of Coefficients

Before proceeding into a discussion of my statistical models, it will be useful to understand how to interpret the results in Table 4.1. What the intercept represents in this MEM is crucial to understanding the quantitative results presented in this paper. This coefficient is the natural logarithm of the predicted RT of an L2 speaker with a personal mean reaction time (PMRT) of 2000 ms for an *n*-gram that a) had the lowest possible MI in the sample, b) had the lowest possible frequency in the sample, and c) was the first stimulus item presented. The value of 2000 ms for a PMRT was selected because it was the mean RT for L2 speakers in this sample. The intercept represents this baseline group and was used in combination with the other coefficients to compute expected RTs for participants who fell into different categories (e.g., an L2 speaker with a 1500 ms PMRT, judging a medium MI, high frequency item, as the final stimulus presented). With this explanation in mind, I will describe my model selection process in the following sections.

4.1.2 Model A

Model A corresponds to Singer & Willet's (2003) concept of an unconditional growth model or Kreft & de Leeuw's (1998) random coefficient model. The only predictors included were the level-1 predictors of MI, frequency, and ordering. Only MI and frequency were allowed to vary, as the ordering variable is not a predictor of interest in this experiment but rather a covariate that needs to be partialed out of the equation. No level-2 predictors were included in this model. Significant effects for frequency and ordering were found.

4.1.3 Model B

Model B attempts to represent Ellis et al.'s (2008) statistical analysis within the framework of an MEM. In this model, RT_{log} was regressed on the level-1 predictors of MI and frequency, with the level-2 predictor of speaker status predicting variation in the level-1 coefficients for the intercept, MI, and frequency. Ordering was not included in this model.

Significant effects for frequency, speaker status on the intercept, and speaker status on the MI-slope term were found. This model suggests that L1 speaker's baseline RT is significantly faster than L2 speakers, that L2 speakers process more frequent phrases more quickly, and that L1 speakers process phrases with a greater degree of association significantly more quickly than L2 speakers. The lack of a significant interaction between frequency and L1 speaker status does not indicate that L1 speakers are insensitive to the effects of frequency – rather, that the effect of frequency for L1 speakers is not significantly different from that of L2 speakers.

4.1.4 Model C

Model C included the same level-1 predictors as Model B; however, the random effect structure was changed. The intercept term was predicted by personal mean reaction time (PMRT), while the MI and frequency coefficient were predicted by L1 speaker status. Significant effects were found for PMRT on the intercept: Slower participants have a slower predicted baseline RT. A significant effect was also found for frequency: L2 speakers judge more frequent phrases more quickly. A significant effect of MI was found for L1 speakers relative to L2 speakers: L1 speakers responded more quickly to more strongly associated phrases. There was no significant effect of MI found for L2 speakers. No effect for frequency was found for L1 speakers beyond that which was present for L2 speakers: That is, L1 speakers process more frequent phrases more quickly; however, this effect is not significantly different from the effect found for L2 speakers.

4.1.5 Model D

Model D is the most theoretically and statistically appealing model tested. The model specifications are much the same as Model C; however, since the effect of frequency was not significantly different across speaker groups, the effect of frequency was fixed so as to reduce the number of parameters estimated by the model. This gives a parsimonious model of theory that is more in line with what the data show than the others examined. Although fixing the effects of frequency results in a slight uptick in model deviance, the corresponding AIC and BIC goodness-of-fit indices both decrease, indicating that this model fits the data better than Model C. In the absence of a significant Analysis of Deviance test, the lower AIC and BIC provided statistical evidence for settling on this theoretical model.

Below are two mathematical representation of Model D. *RT* corresponds to the natural logarithm of reaction time, *MI* to mutual information score, *Freq* to frequency, and *MRT* to personal mean RT (centered at 2200 ms, the mean RT of L2 speakers).

$$RT_{ij} \sim \pi_0 + \pi_1 MI_j + \pi_2 Freq_j + \pi_3 Order_j + \epsilon_{ij} \tag{4.1}$$

$$\pi_{0i} \sim \gamma_{00} + \gamma_{01} Status_i + \gamma_{02} MRT_i + \zeta_{0i} \tag{4.2}$$

$$\pi_{1i} \sim \gamma_{10} + \gamma_{11} Status_i + \zeta_{1i} \tag{4.3}$$

$$\pi_{2i} \sim \gamma_{20} \tag{4.4}$$

$$\pi_{3i} \sim \gamma_{30} \tag{4.5}$$

For reference, the final composite model follows:

$$\begin{aligned} RT_{ij} \sim & \gamma_{00} + \gamma_{01} StatusL1_i + \gamma_{02} MRT_i + \gamma_{10} MI_j + \gamma_{11} StatusL1_i MI_j + \\ & \gamma_{20} Freq_j + \gamma_{30} Order_j + \\ & \zeta_{0i} + \zeta_{1i} MI_j + \epsilon_{ij} \end{aligned} \tag{4.6}$$

4.2 RESULTS OF MODEL D

The substantive results are similar in many respects to Ellis et al.'s and Ellis & Simpson-Vlach's original findings. To summarize the results of Model D: First, those who are faster in general have a faster baseline reaction time. This is likely in part due to L1 speakers having had more exposure to English, with this increase in exposure resulting in lower processing times. It is important to note, though, that there is overlap in the ranges of the PMRTs (cf. Table 3.7, p. 46) of the two groups—some L2 speakers are faster than some L1 speakers. This could be a function of these L2 speakers achieving higher levels of English language proficiency in academic vocabulary and formulaic sequences than some of L1 speakers in this sample. Furthermore, other factors such as age, visual acuity or speed of visual processing may have exerted an effect on exactly how fast a participant was. These factors were beyond the scope of this study but further examination of these effects would be useful.

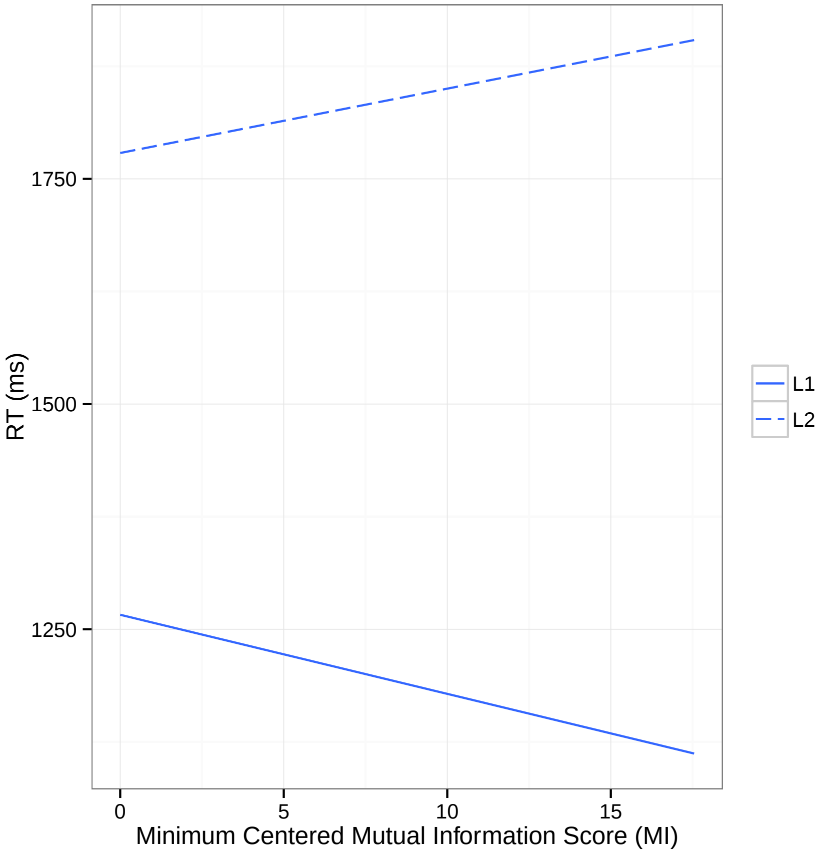
Second, the Level-1 covariate of order of stimulus presentation was also significant, indicating that participants became quicker at responding to the test items as the experiment progressed. Since this covariate is not of particular substantive interest in this study, I will note its statistical significance and move on.

Third, the significant cross-level interaction between L1 speaker-status and MI showed that L1 speakers process more strongly associated n-grams more quickly than L2 speakers do. In contrast, the non-significant increase in RT for higher degrees of MI for L2 speakers indicates that the effect is not present in this sample or is unable to be estimated with confidence. Figure 4.1 illustrates this effect. The x-axis corresponds to the MI of each of the 81 stimuli, while the y-axis represents RT as calculated from the coefficients in Model D. The lines in the graph illustrate the cross-level interaction of MI and Speaker group. As such, the differences in the slopes of the lines is of primary interest, not the magnitude of their separation. The negative slope of the L1 speaker group shows that L1 speakers tended to judge highly associated n-grams more quickly than less associated ones. The slight positive slope of the L2 speaker line, conversely, indicates a non-significant increase in RT for each point of MI.

Fourth, L2 speakers process more frequent n-grams more quickly, irrespective of the phrase's MI. The non-significant effect of the cross-level interaction of L1 speaker status and n-gram frequency must be interpreted with caution. It does not indicate that L1 speakers do not have some processing advantage for phrases of greater frequency, but rather that their sensitivity is no stronger than that measured for L2 speakers. The lines in the Figure 4.2, which are for all intents and purposes parallel to one another, illustrate this. The x-axis corresponds to the frequency of each of the 81 stimuli, while the y-axis represents RT as calculated from the coefficients in Model D. The parallel lines in the graph illustrate the lack of a cross-level interaction of frequency and speaker group. Again, the differences in the slopes of the lines is of primary interest, not the magnitude of their separation. The negative slopes of both groups' lines show that both L1 and L2 speakers tended to judge more frequent n-grams more quickly than less frequent ones.

Thus far, the results broadly support Ellis and colleagues' findings—namely,

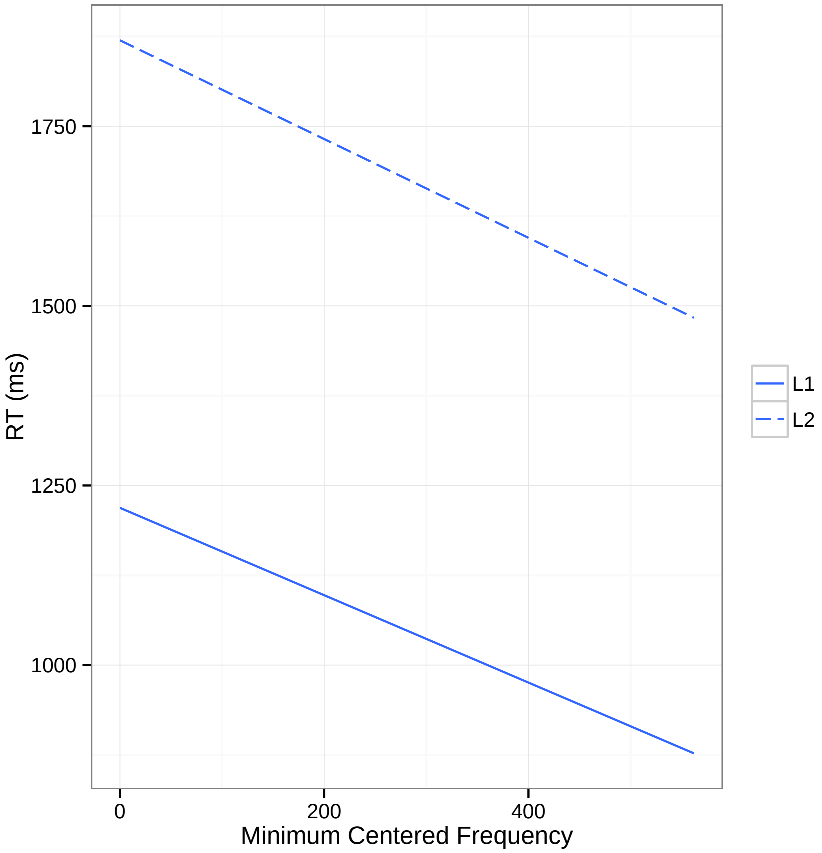
Figure 4.1: Fitted Values vs. Mutual Information



that L1 speakers process more associated n-grams more quickly and that L2 speakers process more frequent n-grams more quickly. However, L1 speakers in this sampled showed a similar degree of sensitivity to frequency as L2 speakers. This latter finding different from Ellis et al.'s studies that did not find this frequency effect for L1 speakers.

However, some evidence exists in these data that indicates not all L1 participants have developed this MI sensitivity and that some L2 speakers have. Compare participants 101, 109, and 112 in the Figure 4.3 to others (e.g., 107 or 108). Although the former participants are L1 speakers of English, there is a definite positive slope for the lines-of-best-fit in their scatter plots, indicating that they

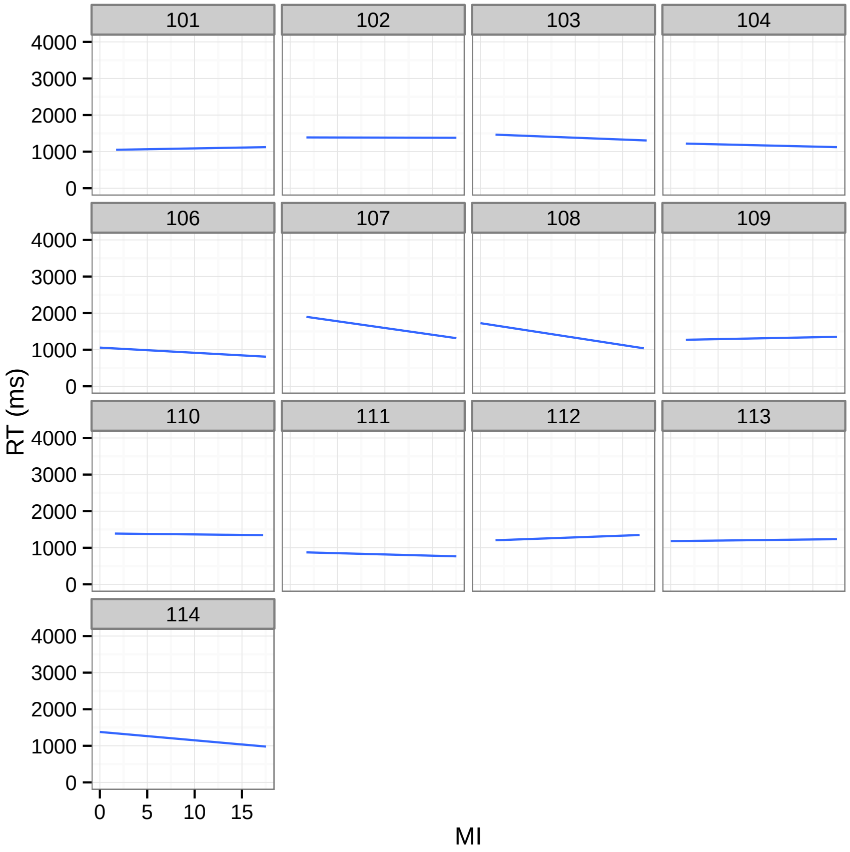
Figure 4.2: Fitted Values vs. Frequency



did not process more highly associated *n*-gram more quickly. I have no clear hypothesis as to why this is because of the variability in their responses to how much academic reading each participant does per week (1 hour, 15 hours, and 14 hours) respectively. Two of the participants had recently completed master's programs after 2.5 years, while one had only been in their master's program for 6 months.

Conversely, these data show it may be possible for this sensitivity to develop in highly proficient L2 speakers because some L2 speakers did respond more quickly to more highly associated items. Consider participants 202, 208, and 210 in Figure 4.4 in relation to participants 204 and 209. The first group of participants were among the most educated in the group, with one having earned a PhD, one currently working on a PhD, and one finishing a master's degree. Participants 202,

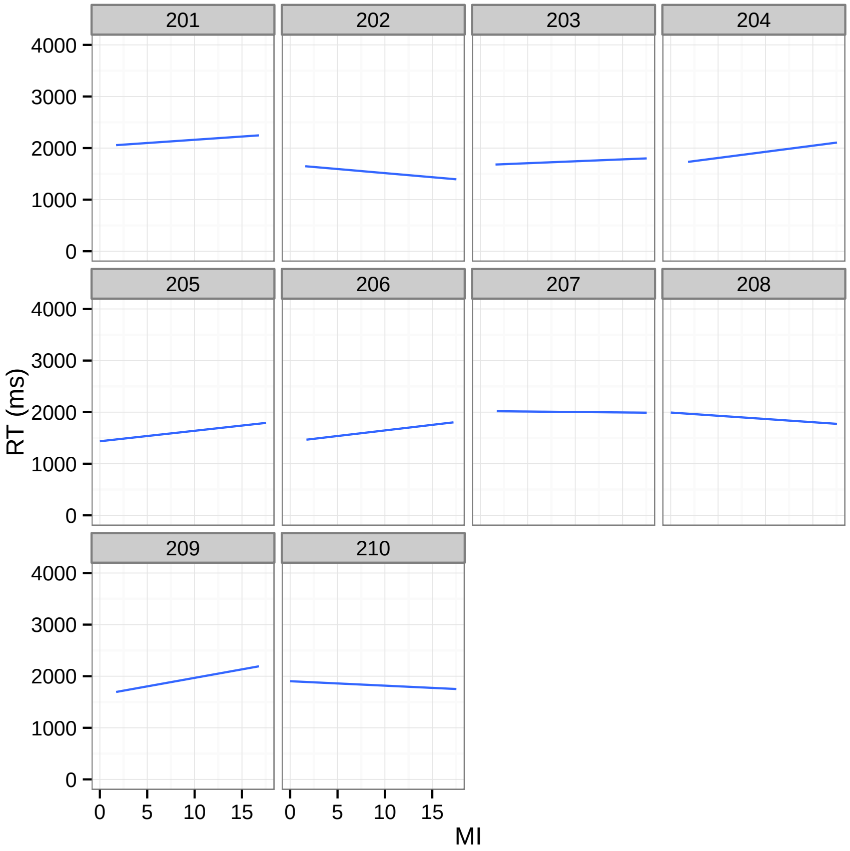
Figure 4.3: Individual Fitted RT Values vs. MI for L1 Speakers



207, and 210 had all lived in a primarily English-speaking environment for more than 5 years. Living in a primarily English-speaking environment for many years may provide the L2 user with much more opportunity to process English in non-academic contexts. It may be that these speakers have processed enough individual words and phrases of English to have a better implicit sense of just how unusual it is for the combinations of words that make up highly frequent, associated *n*-grams to occur together.

The results arising from the analysis of the individual participant trends offer a more detailed picture of the phenomenon under investigation and illustrate possible future avenues of investigation—e.g., at what rate do these sensitivities develop? In the next chapter, I shall examine the implications of these results and consider

Figure 4.4: Individual Fitted RT Values vs. MI for L2 Speakers



directions for future research.

Chapter 5

CHAPTER 5

In this chapter, I shall discuss the results of my study. First, I will recap the findings of my analysis from the previous chapter and provide the results of a quantitative thought exercise in order to the possible real-world implications of the results. Second, I will connect the results of my study to emergentist models of language. Third, I will explore some of the limitations of this study. Finally, I will discuss the implications of this study for TESOL, psycholinguistics, the quantitative side of linguistic research, and neurolinguistics in the context of language pathology and development.

5.1 A THOUGHT EXPERIMENT IN PRACTICAL SIGNIFICANCE

In this section, I will restate and expand on the practical significance of the results detailed in Chapter 4. Using values predicted by model for different types of speakers judging *n*-grams of different frequencies and associational levels, I will show how the average L2 speaker may experience a 7 minute increase in processing time over the course of reading a 10,000 word academic article.

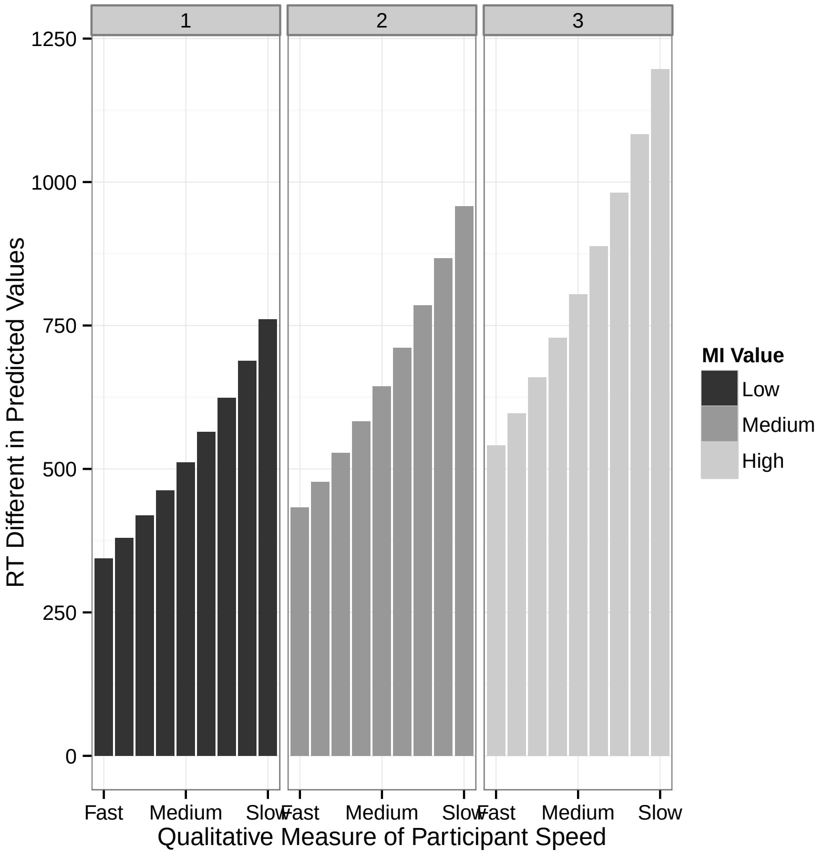
As I showed in Chapter 4, MI was a significant predictor of reduced reaction time in L1 speakers of English. Frequency, on the other hand, was a significant predictor of reduced reaction time for both L1 and L2 speakers of English. These results stand in contrast to Ellis and colleagues' findings, which were that MI alone was the predictor associated with reduced reaction time for L1 speakers, while frequency was predictive of reduced reaction time only in L2 speakers. These results raise an important question: What do they mean in real world terms? Using the calculations of Biber, Johannson, Leech, Conrad & Finegan (1999, pp. 990-1000), 3-word frequent, associated n-grams occur approximately 60,000/million words in academic writing, while similar 4-word n-grams occur at a rate of approximately 5,000/million words. Assuming that the findings of this study can be generalized to 3-word n-grams in academic writing, how much extra time might L2 speakers spend processing these 3- and 4-word sequences?

In order to further investigate the practical significance, I compared the predicted—fitted—values for an L1 of average PMRT for L1 speakers, and an L2 speaker of average PMRT for L2 speakers. These calculation assumed that the n-gram to process is of average frequency, average MI, and average order in the presentation of stimuli. Using Biber et al's (1999) figure of 65,000 3- and 4-word n-grams per million words, this results in a 630.1 ms difference in processing time for *each* n-gram between the average L1 and L2 speaker processing the average n-gram. This translates to approximately 683 extra minutes for reading the equivalent of 10 mass-market paperback books of 300 pages (Biber et al., 1999). For a 10,000 word article this is an extra 6.83 minutes of processing time for these types of n-grams *without* regard for other syntactic processing, semantic processing, age effects, or content comprehension. Note that this estimate does not take into account the difference in processing between a very fast L1 speaker and an average or slower-than-average L2 speaker.

Figure 5.1 below elucidates the situation further. The x-axis represents qualitative measurement of speaker speed relative to their group. Since L1 and L2 speakers have different, but overlapping, ranges of values for PMRTs, I rank-ordered the possible RTs, so that, for example, the fastest L1 and L2 speakers can be easily compared on the graph. Fast corresponds to those speakers who reacted very quickly to the stimuli relative to the rest of their group; Slower corresponds to those who reacted more slowly. The y-axis represents the difference in RTs predicted for

L1 and L2 speakers of the same relative speed, while processing n-grams of average frequency, average rank in the order of presentation across different levels of mutual information. The different panels correspond to different levels of MI. This figure shows that the processing gap is greater in those with slower PMRT and that MI increases, the processing gap between L1 and L2 speakers widens. Furthermore, this gap is even more pronounced among slower reactors in both groups—this is to say that the magnitude of the difference is greater on the right side of the figure.

Figure 5.1: Differences between L1 and L2 speakers in Predicted RTs



5.2 IMPLICATIONS: CONNECTION TO THEORY

In this section, I shall consider this study's results in relation to Emergentist models of language. These results, although not an exact replication of Ellis et al.'s, tell a similar story: L1 speakers are sensitive to the relative frequency of co-occurrence of the words in multiword phrases. These results suggest that L1 speakers have extra processing advantage beyond the frequency of a phrase that even many highly proficient L2 speakers lack. However, in this sample, both L1 and L2 speakers showed sensitivity to frequency. There is also evidence that highly proficient L2 speakers have developed this sensitivity to MI and that some L1 speakers have yet to do so. These results are interesting in light of the studies that have found reduced processing speed to be a function of increased frequency in L1 speakers (Arnon & Snider, 2010; Bannard & Matthews, 2008; Durrant & Doherty, 2010).

The findings of this study may be evidence for Langacker's (2000) idea of *entrenchment*. Langacker proposed that the human brain forms webs of neural structures that activate in very complex patterns. He suggested that after repeated exposure to a stimulus, such as a word or phrase in context, a particular web, corresponding to a unique set of identifiers that sets apart the word or phrase, would activate. With enough exposures, the web activation leaves traces. When the whole web comes to be entrenched, the traces have accumulated such that the web acts as a whole unit. Sensitivity to either MI or frequency may be indicative of different levels of entrenchment. Sensitivity to frequency may be the first step towards sensitivity to MI. This is to say, MI could be an advanced case of frequency sensitive, as MI sensitivity may be an implicit awareness of the the relative co-occurrence of lexemes. Given the results of examining the individual distributions of RTs, it may be that MI sensitivity is not a marker of the so-called "native speaker", but rather a high degree of proficiency with a particular register. High MI (e.g. *on the other hand*) phrases may be more salient than less associated (e.g. *than that of the*) ones to those with a great deal of proficiency in a given language or register (Durrant & Schmitt, 2009).

Previous research has shown that L2 users tend to underuse highly associated, low-frequency collocations (e.g., Durrant & Schmitt, 2009). Results like Ellis et al.'s and my own can help explain why: Although these phrases occur much more often than chance alone would predict (given the frequencies of the constituent words), they simply do not occur that often! Thus, L2 speakers, who seem more responsive to frequency, may not actually see these enough to form the type of strong neural pathways that arise from exposure to high frequency n-grams. Thus the salience of these phrases is perhaps less apparent to L2 speakers. L2 speakers are unlikely to have had levels of exposure to English typical of those who began acquiring the language in infancy and have had their schooling in English from a young age. Thus, their implicit tallies of the frequencies of lexical units, be they orthographically mono- or polylexical, and the collocations of the the lexical units (Ellis, 2002) are not enough to make it salient that a phrase occurs more often than chance alone would predict. This is to say that L2 speakers have not enough language exposure to recognize the MI of a phrase.

5.3 LIMITATIONS

The results detailed above provide some insight into the human processing of frequent, associated n-grams; however, this study does has several limitations related to the sample, the stimulus generation process, and the measure of MI. First, the sample of convenience, containing only 23 participants, was a small. In this case, more participants would increase power for the MEM more so than the addition of additional stimuli (Kreft & de Leeuw, 1998). Ideally, future studies in this area will include a larger, more heterogeneous sample.

Second, the sample itself was unbalanced. In the sample was a severe gender imbalance in L1 speaker group (11 women, 2 men). Moreover, Applied Linguistics and TESOL were over-represented in the L1 speaker group. Additionally, the L1 group contained two participants who were much older than most other participants. I reran the analysis, excluding the judgments made by those participants. The recalculation of effects did not change the significance of the fixed effects. This suggests that the results were not unduly influenced by the presence of more experienced scholars in the sample. Future inquiries along these lines should endeavor to recruit a more balanced sample.

Third, the selection of stimuli was not as stringent as other studies investigating elements of this phenomenon. Firstly, I did not attempt to control for the frequency of the subgrams (i.e., the constituent n-grams and non-contiguous collocates that make up an n-gram of greater than 2 words) nor degree of association between those elements. Shaoul & Westbury (2011) noted that the re-analysis of Arnon & Snider's (2010) data found that fourth word frequency may contribute to faster processing time as well. Given these results, it is possible that frequencies and associations of the constituent parts of n-grams may also play a role in which phrases are most associated with reduced processing time. Controlling for these factors would allow researchers to understand better which parts of a phrase contribute to its processing. Secondly, I did not attempt to isolate n-grams associated only with academic writing (Ellis & Simpson-Vlach, 2009; Ellis et al., 2008). However, my results are similar to these studies in spite of this relatively basic selection process. Furthermore, the results also show an effect for L1 speaker frequency similar to the studies conducted by Bannard & Matthews (2008) and Arnon & Snider (2010).

Fourth, this study was conducted in a highly experimental setting. Although phrasal decision tasks like this involve larger stretches of language than lexical decision tasks, the nature of the experiment focuses the participant's attention on the language, rather than propositions that the language encodes (Wray, 1992). Participants viewed the stimuli outside of a natural or naturalistic setting, so any generalizations arising from this study applied to situations outside this experimental setting must be viewed with caution. Although the above thought exercise that attempts to make the statistical results relevant to the real world provides a possible consequence of this slower reaction time, to what extent is it really generalizable? When it is truly a matter of human-imperceptible milliseconds, perhaps this 7 minute processing differential estimate is somewhat accurate. However, when a reader encounters n-grams that are very difficult to comprehend (especially for L2 speakers), is it realistic to assume that a reader really continue trying to comprehend a difficult-to-understand phrase it in a very large body of text or just move on (Alison Wray, personal communication)? Future opportunities for study in sort of setting could involve corpus-derived self-paced reading tests, so as to more closely approximate the way that English language users would typically encounter the phrases.

Finally, it is unclear how the calculation of MI for multiword sequences was implemented in Collocate (Barlow, 2010). Given the many frequencies and associations that exist for the components of n-grams, it is unclear what MI measures exactly. Although Church & Hanks's (1990) formula for continuous and non-contiguous bigrams is often cited as the canonical way of determining MI, I have been unable to find definitive methods of calculating it for n-grams of 3 words or more. In the future, I will use a different collocation extraction method that allows for examination of the source code. The open-source model would allow for me to examine the calculation of MI or implement my own.

5.4 IMPLICATIONS FOR LINGUISTICS

In spite of the limitations of this study, the results have implications for TESOL, psycholinguistics, the quantitative, and the analysis of linguistic data. In this section, I shall discuss what the results of this study imply for these linguistic subfields.

5.4.1 TESOL

The results of this study may be useful information for ESL instructors and materials developers. As noted above, some studies have shown that even advanced English learners may underuse highly associated, relatively infrequent collocations (Durrant & Schmitt, 2009), while others have shown that L2 speakers of English may not have as finely developed sense of collocational frequency (Siyanova & Schmitt, 2008). These previous results in combination with my results and the results of Ellis et al. (2008) suggest that L2 learners may need extra exposure to these infrequent, but associated collocations and *n*-grams.

Although the reduction in processing of frequent, highly associated *n*-grams time that L2 speakers may experience is small amidst a sea of orthographic, syntactic, semantic, discursive, and visual processing necessary to read a text, this is an area that can be addressed through pedagogical intervention. In order to address this potential need, a trajectory for development phrasal learning materials for L2 academic writers of English might focus on two axes of formulaicity: frequency and association. Instead of only choosing the most frequent *n*-grams, teachers or other materials developers could look for lower frequency phrases that are *also* highly associated, as Ellis et al.'s research found a correlation between the evaluation of pedagogical usefulness by ESL professionals and degree of phrasal association. Simpson-Vlach & Ellis's (2010) Academic Formulas List (AFL) attempts to do just that, as a generalization of Coxhead's (2000) Academic Word List to the polylexical case. Ellis et al. constructed a list to include the most pedagogically relevant academic multiword sequences, as observed from the analysis of academic spoken and written corpora. Similarly, Martinez & Schmitt (2012) developed a phrasal expression list that focused on the most frequent, semantically opaque phrases in the British National Corpus. Conrad & Biber (2009) and Peake (2011) also attempted to develop materials for advanced learners of English through the principled examination of corpora and the inclusion of phrasal items.

5.4.2 Psycholinguistics

In addition to the ramifications for TESOL, my study shows the viability of the use of corpora to generate psycholinguistic research. For researchers of formulaic language, corpora and concordance tools are especially useful for generating stimuli that depends on quantitative measures (e.g., collocations, lexical bundles, or *n*-grams). Although corpus extraction based on frequency and degree of association will not include all types of FL (Wray, 2008), it provides a useful way to obtain stimuli with certain properties that clue us into the possibility that the phrases are, in fact, formulaic. Whether corpora are part of the stimulus generation process (Bannard & Matthews, 2008; Ellis & Simpson-Vlach, 2009; Ellis et al., 2008; Siyanova & Schmitt, 2008) or as method to verify or challenge one's assumptions about proposed stimuli (Conklin & Schmitt, 2008), the use of corpora makes it a simple matter for researchers to find test items that are examples of actual, attested language or to consider assumptions about stimuli empirically. These corpus-based approaches present an interesting opportunity to study the processing of language that is commonly used. It is no longer necessary to search out the dusty corners of our linguistic imaginations to test the boundaries of language or rely entirely on our intuitions about 'what is formulaic' or 'what sounds right'.

5.5 QUANTITATIVE IMPLICATIONS FOR LINGUISTICS

The results of this research have implications for quantitatively oriented researchers in applied linguistics as well as quantitative psychologists and statisticians. In this section, I will discuss the usefulness of MEMs for applied linguists and explain briefly how the results of this study could be useful in power analyses of a MEM used to study this phenomenon.

First, as I argued above, MEMs address many of the problems that are associated with historical methods of treating dependency in OLS regression models.

MEMs are a more accurate and elegant method than aggregate OLS regression analysis for examining intra- and interindividual differences in a single model. Chiefly, MEM eliminates α-inflation that arises from running multiple tests. Ellis et al. (2008) employed two aggregate OLS regression models. However, the researchers did not note any correction for running multiple tests, which is a source of α-inflation—i.e., the probability of making a Type I error increases. Secondly, MEMs allow for the unbiased estimation of fixed effect parameters and their standard errors (Kreft & de Leeuw, 1998). Ellis et al. (2008) and Ellis & Simpson-Vlach (2009), employed aggregate OLS regressions model in their analyses. This may be a relatively conservative approach because of the possibility that the parameter standard error estimates are inflated (Cohen et al., 2003), as over-estimated standard errors lead to lower values of test statistics, and thus higher p-values. Although this approach may be better than simply not attempting to account for dependency, the deflation of test statistics does not tell us the most accurate story. A MEM might allow for more precise standard error estimation and might detect significant effects (that were only marginally significant before), such as in Ellis & Simpson-Vlach's (2009) Experiment 4, where MI was found only to approach significance as a predictor of reduced RT. Finally, MEMs do not sacrifice the individual distributions of the dependent variable, so it is possible to look more closely at the data, as I did above to identify possible trends and explanations for patterns seen in the data.

Because of the closer look into one's data allowed through mixed-effect modeling, MEMs are a highly relevant technique for applied linguists beyond the domain of experimental psycholinguistic studies. For instance, MEMs have been used in educational psychology to investigate situations where students are nested within different classes (Singer & Willet, 2003). However, these studies have focused on K-12 education. English language education programs, such as PSU's IELP, could also benefit from the application of MEMs to larger scale studies of their program.

These models are not just tools for psycholinguists or sociolinguists, but for anyone in applied linguistics collecting nested data. The greater precision offered by MEMs lends a greater degree of confidence to the accuracy and validity of the results.

Finally, this research has produced parameter estimates that could be used as a starting point for power analysis of MEMs for research concerned with the effects of frequency and association on reaction time. Power analysis serves as means for detecting a model's Type II error rate and may be employed to estimate sample size needed to detect effects of the expected size. For MEMs this process can be difficult (Cohen, Cohen, West, & Aiken, 2003; Kreft & de Leeuw, 1998). Unlike many familiar univariate techniques such as *t*-tests, ANOVA, multivarible regression, there are not easy closed-form solutions to the ultimate power of a MEM, so these calculations are often carried out via simulation studies (Kreft & de Leeuw, 1998). Since these analyses need approximations to start the iterative calculations, the results provided above can be used as effect estimates.

5.6 DIRECTIONS FOR FUTURE RESEARCH

In this section I will suggest some directions for future research. First, I will consider some of the issues surrounding the calculation of MI and propose an extension of this calculation to bound morphology. Second, I will discuss some directions for future inquiries into the processing of frequent, associated *n*-grams. Finally, I will conclude with some thoughts on the study of language development and pathology from a neurolinguistic perspective.

5.6.1 Extended MI and Applications

As I showed in Chapter 2, the calculation of MI for multiword sequences is beset with potential complications. Without knowing how MI is being computed, it is unclear whether the reduction in processing time that highly proficient English users experienced is due to the association of the whole phrase or perhaps some lower level phrasal or lexical component. Although Church & Hanks's (1990) formula for continuous and non-contiguous bigrams is often cited as the canonical way of determining MI, I have been unable to find definitive methods of calculating it for n-grams of 3 words or more. Many of the closed-sourced, proprietary software packages do not elucidate how MI is calculated. Given the various ways one could calculate multiword MI (Morgan, 2014; Trnka, 2011), the active exploration of what these measurements actually mean is necessary. For instance, how does MI behave in a corpus when different subphrase frequencies are taken into account? Could it be the frequencies and associations of the constituents parts that are more predictive of a reduction in processing time (Shaoul & Westbury, 2011)?

5.6.2 Psycholinguistics

Within the field of psycholinguistics, there are many opportunities for further scholarship to examine the the processing of associated elements of language. In this section I will discuss several other questions that have arisen from this research.

First, will English language users with the highest proficiency in a register be sensitive to degree of association? Several possible follow-up studies could be conducted to investigate this. For instance, a methodologically similar study could be conducted consisting of four groups: professional academic writers, advanced graduate students, beginning graduate students, and undergraduates. Viewed in the light of a proficiency-based sensitivity to MI, it follows that the most proficient language users, professional academics and advanced graduate students would exhibit this sensitivity to phrasal association, whereas the other two groups would not. If such a study included L2 speakers, comparisons could be made across groups as to further test the proficiency hypothesis.

Second, another possibility would be to investigate this phenomenon in an entirely different register, such as conversation or newspaper writing. The newspaper genre provides an particularly interesting opportunity for research. Given the massive transition to digital media and news sources, as evidenced by *The Oregonian's* recent service reduction to four days a week, there is a divide in where older and younger language user obtain news. Younger speakers are more likely to get their news online from a variety of sources, while many older speakers may prefer to obtain news in print form. Two questions arise out of this: Are there differences in the use highly frequent, highly associated n-grams in digital and print news? If so, how do younger and older speakers process the n-grams most associated with each? By looking outside of academia, researchers interested in this aspect of FLA may be able to find a larger pool of participants.

Third, studies on this topic could examine L1 and L2 speakers either with truly longitudinal means or a cross-sectional design with speakers of different different age groups. A non-linear Mixed Effects Model (MEM) exploration of this relation would be an interesting avenue to explore because it could show when sensitivity to frequency and MI develop and whether frequency sensitivity is a precursor to MI sensitivity.

Finally, a further avenue for exploration would be to consider would be to consider what extent the production of frequent, highly associate$n$-grams is correlated with the reduction in processing time experienced by proficient English users. Since reception and production of language seem to result to some degree in the activation of similar neural structures. It is possible that to become sensitive to MI, one must also produce the phrases in context.

This might explain the difference between the L1 speakers who did not show the processing advantage typical of their group and the L2 speakers who did. Three of the L2 speakers who showed this sensitivity had lived in the United States for more than 5 five years and had spent at least four years in graduate school or as an academic professional. The other participant had only lived in the United States for 1.5 years but was studying at the PhD. level, in Engineering wherein much technical writing is done in English. It is likely that all of these participants had more chances to practice academic writing than the average participant. This practice may have been another source of neurological activation resulting in the entrenchment of the phrases and thus a processing advantage for association. In contrast, the L1 speakers who showed an atypical pattern of MI sensitivity relative to their group all had graduate education but none more than 2.5 years. Perhaps their sensitivity to frequency but not MI was not due to a lack of exposure, but a lack of production?

5.6.3 Neurolinguistics and Clinical Linguistics

The findings of this study have also raised questions for further neurolinguistic research in language pathology and development. In this section, I will examine some reasons why linguists might use neuroimaging techniques to examine the neurological processing locus of highly associated *n*-grams. In this section, I will explain how a better understanding of associated *n*-gram processing broader implications and directions for future research that have arisen from this research.

Further studies from a neurolinguistic perspective may elucidate the physical processes associated with the processing and storage of formulaic language. Since there seems to be a psychological processing advantage for frequent, highly associated *n*-grams, the question naturally arises: What are the neurological structures associated with increased activation during such processing. Examining this phenomenon through EEG and fMRI techniques could be useful in understand the effects brain damage to either hemisphere. Through this process, it would be possible to learn about the neurological lateralization of processing highly frequent, highly associated *n*-grams. If the processing and production of these phrases is associated with one particular hemisphere, it might explain why one person with brain damage might be unable to use collocations; whereas for a patient with a different pattern of brain damage, collocations might be an exploitable resource for meaning-making. Is the processing and production of highly associated n-grams more associated with the right or left hemisphere?

Several studies in studies of language pathology have provided insights into the neurological loci of formulaic language (Sidtis, Canterucci & Katsnelson, 2009; Van Lancker Sidtis, 2006; Van Lancker Sidtis & Postman, 2006). In short, patients with left-hemisphere brain damage used a greater proportion of formulaic language, while those with right-hemisphere damage used less formulaic language and experienced difficulty in interpreting non-literal meanings (among other things). While these studies suggest that have shown that some types of formulaic language (e.g. idioms, proverbs) are associated with elevated neural activity in in the right hemisphere of the brain or basal ganglia (e.g. swearing, serial speech), I am not currently aware of any studies examining the processing of the n-grams in this clinical context. Thus, several questions follow. For instance, to what extent do people with neurological damage use collocations? If collocations and associated n-grams are included in studies as part of the definition of formulaicity, do right-hemisphere damaged patients really use a smaller proportion of formulaic language? To what extent are collocations and associated n-grams retained in different language pathologies? If some or all of these sequences are stored in the right hemisphere, it is possible that they would be preserved longer in those who suffer right-hemisphere damage. Then it follows, that they might be less present in left-hemisphere damaged patients. If this were found to be the case, it would imply that the mechanism for frequency sensitivity reside in the right hemisphere. Conversely, if these phrases are mostly left-hemisphere lateralized, it follows that the sensitivity to relative frequency of lexemes is left-hemisphere lateralized. Examining the using and processing of n -grams in cases of language pathology, could shed light on how these are processed and help cognitive scientists better understand where the human sensitivity to frequency resides.

In addition to situations of language pathology, neurolinguistic studies of this nature might elucidate the processes of acquiring highly associated *n*-grams. For instance, what areas of the brain show activation in younger learners when they are exposed to these phrases? Are speakers processing these items with the right hemisphere, thus perhaps indicating more holistic processing? Conversely might it be that speakers are engaged in left-hemisphere processing and are engaged in a more decoding type of processing? To what extent are *n*-grams retained through the life span? Does the processing advantage imparted by frequency or MI remain without relatively frequent activation? An interesting follow up study could examine processing of these phrases by current and retired professional academic writers. By controlling for age and time since leaving academia, it would be possible to learn more about the processing and retention of these phrases over the whole human lifespan.

5.7 CONCLUSION

In conclusion, this study has replicated one of the most interesting findings of Ellis et al.'s original research: that L1 speakers process highly associated *n*-grams more quickly and that L2 speakers of English lack this sensitivity. My study did find a significant effect of frequency for both L1 and L2 speakers, in contrast to Ellis et al. These results have practical, as well as statistical significance: Reading a 10,000 word article may require 7 or more minutes of extra processing time for an L2 English speaker just to process the frequent, highly associated *n*-grams. Although this is a relatively small amount of time, over a million words, it results in an extra 10 hours of processing time. The good news is, though, that this is an area that ESL materials developers and teachers can address. An ESL teacher may have a very difficult time helping a student comprehend the content of the student's substantive area; however, the same ESL teacher could provide opportunities for the learner to see highly-associated and less frequent n-grams in context to ease some of the processing burden. With those cognitive resources freed, the student may be able to focus her attention on other areas of processing difficulty (e.g., syntactic, semantic, or content processing).

In the coming years, it will be important to consider data from FLA, SLA, and clinical linguistics to better understand not only development up to adulthood but through the entire lifespan in order to develop a complete theory of language. It is crucial that LA not be considered "done"—language acquisition, in fact, continues through the entire lifespan in all adults, irrespective of neurotypicality. Craig Finn said that "Certain songs, they get so scratched into our souls." Although this study does not purport to show evidence of language being etched into a soul, it does suggest that some types of language, may be(come) etched into our minds, and by extension, the neurological hardware that generates them.

REFERENCES

Abutalebi, J., Cappa, S. F., & Perani, D. (2001). The bilingual brain as revealed by functional neuroimaging. *Bilingualism: Language and Cognition*, *4*, 179–190.

Ädel, A. & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, *31*, 81–92.

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

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.

Barlow, M. (2010). *Collocate.* Houston, TX: Athelstan. Software for collocation extraction.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4.* R package version 1.1-6.

Biber, D. (2006). Stance in spoken and written university registers. *Journal of English for Academic Purposes*, *5*, 97–116.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*, 371–405.

Biber, D., Johannson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman grammar of spoken and written English.* London: Longman.

Bloor, T. & Bloor, M. (2004). *The functional analysis of English.* London: Hodder Arnold.

Bridges, K. A. & Van Lancker Sidtis, D. (2013). Formulaic language in Alzheimer's Disease. *Aphasiology*, *27*, 799–810.

Chafe, W. (1968). Idiomaticity as an anomoly in the Chomskyan paradigm. *Foundations of Language*, *4*, 109–127.

Chen, Y.-H. & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology*, *16*, 30–49.

Church, K. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*, 22–29.

Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences.* New York: Routledge.

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

Conrad, S. & Biber, D. (2009). *Real Grammar: A Corpus-Based Approach to English.* London: Pearson Education ESL.

Coxhead, A. (2000). A new Academic Word List. *TESOL Quarterly*, *34*, 213–238.

Davies, M. (2008). The Corpus of Contemporary American English: 450 million words, 1990-present. http://corpus.byu.edu/coca.

Dunning, T. (1993). Accurate methods for the stastistics of surprise and coincidence. *Computational Linguistics*, *19*, 61–74.

Durrant, P. & Doherty, A. (2010). Are high-frequency collocations psychological real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory*, *6*, 122–155.

Durrant, P. & Schmitt, N. (2009). To what extend do native and non-native writers make use of collocations. *International Review of Applied Linguistics*, *47*, 157–177.

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

Ellis, N. C. (2001). Memory for language. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 33–68). Cambridge: Cambridge University Press.

Ellis, N. C. (2002). Frequency effects in language processing: A review with impilcations for theories of implicit and explicit language acquisition.

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

Ellis, N. C. & Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory*, *5*, 61–78.

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

Elman, J. L., Bates, E. A., Johnson, M. H., Karmilloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking Innateness: A connectionist perspective of development*. Cambridge, Massachusetts: The MIT Press.

Erman, B. & Warren, B. (2000). The idiom principle and the open-choice principle. *Text*, *20*, 29–62.

Eskildsen, S. (2009). Constructing another language: Usage-based linguistics in second language acquisition. *Applied Linguistics*, *30*, 335–357.

Eskildsen, S. (2012). L2 negation constructions at work. *Language Learning*, *62*, 335–357.

Grothendieck, G. (2014). *sqldf: Perform SQL Selects on R Data Frames*. R package version 0.4-7.1.

Højsgaard, S., Halekoh, U., Robison-Cox, J., Wright, K., & Leidi, A. A. (2013). *doBy: Groupwise summary statistics, LSmeans, general linear contrasts, various utilities*. R package version 4.5-10.

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

Kreft, I. G. G. & de Leeuw, J. (1998). *Introducing Multilevel Modeling*. London: SAGE Publications.

Langacker, R. (2000). A dynamic usage-based model. In M. Barlow & S. Kemmer (Eds.), *Usage-based models of language* (pp. 1–59). Stanford: CLSI.

Larsen-Freeman, D. & Cameron, L. (2008). *Complex systems and applied linguistics*. Oxford: Oxford University Press.

Lindholm, C. & Wray, A. (2011). Proverbs and formulaic sequences in the language of elderly people with dementia. *Dementia*, *10*, 603–623.

MacWhinney, B. (2000). Connectionism and language learning. In M. Barlow & S. Kemmer (Eds.), *Usage-based models of language* (pp. 1–59). Stanford: CLSI.

Martinez, R. & Schmitt, N. (2012). A phrasal expression list. *Applied Linguistics*, *33*, 1–23.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, T. G. B., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182.

Morgan, J. A. (2014). Explorations into the psycholinguistic validity of extended collocations. Paper presented at 2014 Conference of the American Associaton of Applied Linguistics in Portland, Oregon.

Nattinger, J. R. (1980). A lexical phrase grammar for ESL. *TESOL Quarterly*, *14*, 337–344.

Nattinger, J. R. & DeCarrico, J. S. (1992). *Lexical phrases in language teaching.* Oxford: Oxford University Press.

O'Donnell, M., Römer, U., & Ellis, N. C. (2013). The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics*, *18*, 83–108.

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

Peake, B. (2011). Formulaic language for academic writing: Analysis and materials development with text organizing lexical bundles. Unpublished Master's Project.

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*.

Portland State University (2014). English language proficiency requirements: Graduate. Retrieved from http://www.pdx.edu/admissions/english-language-proficiency-requirements-graduate on June 17th, 2014.

Pothos, E. & Juola, P. (2007). Characterizing linguistic structure with mutual information. *British Journal of Psychology*, *98*, 291–304.

R Core Team (2014). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Revelle, W. (2014). *psych: Procedures for Psychological, Psychometric, and Personality Research.* Evanston, Illinois: Northwestern University. R package version 1.4.4.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *Sociological Review*, *15*, 351–357.

Rumelhart, D. E. & McClelland, J. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I.* Cambridge, MA: MIT Press.

Shaoul, C. & Westbury, C. (2011). Formulaic sequences: Do they exist and do they matter? *The Mental Lexicon*, *6*, 171–196.

Sidtis, D., Canterucci, G., & Katsnelson, D. (2009). Effects of neurological damage production of formulaic language. *Clinical Linguistics and Phonetics*, *23*, 270–284.

Simpson-Vlach, R. & Ellis, N. C. (2010). An academic formulas list (AFL). *Applied Linguistics*, *31*, 487–512.

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Singer, J. & Willet, J. (2003). *Applied Longitudinal Data Analysis.* London: Oxford University Press.

Siyanova, A. & Schmitt, N. (2008). L2 learner production and processing of collocations. a multi-study perspective. *The Canadian Modern Language Review*, *64*, 429–458.

Stubbs, M. (1993). British traditions in text analysis: From Firth to Sinclair. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and Technology* (pp. 1–59). Philadelphia: John Benjamins.

Swinney, D. & Cutler, A. (1979). The access and processing of idiomatic expressions. *Verbal Learning and Verbal Behavior*, *18*, 523–534.

Trnka, K. (2011). Pointwise mutual information score. Retrieved on April 14, 2014 from http://acl.ldc.upenn.edu/J/J90/J90-1003.pdf.

Van Lancker Sidtis, D. (2006). Where in the brain is non-literal language? *Metaphor and Symbol*, *21*, 213–244.

Van Lancker Sidtis, D. & Postman, W. (2006). Formulaic expressions in spontaneous speech of left- and right-hemisphere damaged subjects. *Aphasiology*, *20*, 411–426.

Warnes, G. R., Bolker, B., Gorjanc, G., Grothendieck, G., Korosec, A., Lumley, T., MacQueen, D., Magnusson, A., Rogers, J., & others (2014). *gdata: Various R programming tools for data manipulation.* R package version 2.13.3.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis.* Springer New York.

Wray, A. (1992). *The Focusing Hypothesis: The theory of left hemisphere lateralised language re-examined.* Amsterdam: John Benjamins.

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics*, *21*, 463–489.

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

Wray, A. (2008). *Formulaic language: Pushing the boundaries.* Oxford: Oxford University Press.

Wray, A. (2010). 'We've had wonderful, wonderful thing': Formulaic interaction when an expert has dementia. *Dementia*, *10*, 517–534.

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of the field. *Annual Review of Applied Linguistics*, *32*, 231–254.

Wray, A. & Perkins, M. (2000). The functions of formulaic language: An integrated model. *Language and Communication*, *20*, 1–28.

Xu, Y., Jones, G., Li, J., Wang, B., & Sun, C. (2007). A study on mutual information-based feature selection for text categorization. *Journal of Computational Information Systems*, *3*, 1007–1012.

Appendix A

PAIRED STIMULI AND DISTRACTORS

the study of the

the of the study

that it is the

that the is it

to the use of

of to the use

the history of the

the the of history

in the study of

study in of the

to the study of

study of to the

and the number of

number the and of

than that of the

of the that than

of the effects of

effects of the of

to that of the

that the of to

the level of the

the of the level

of some of the

of some the of

the part of the

the of the part

of the use of

use the of of

the form of the

of form the the

the effects of the

the effects the of

the effect of the

of effect the the

the work of the

of work the the

the development of the

development the the of

the structure of the

of the structure the

the results of the

results of the the

the position of the

the the of position

is one of the

the of one is

the use of the

the use the of

in the number of

in the of number

the case of the

the of the case

the time of the

the the time of

is related to the

to the is related

this is not the

is the this not

is similar to that

to similar is that

the first part of

part the first of

to say that the

say that the to

be noted that the

be the that noted

a function of the

the of function a

it is easy to

easy it to is

from the fact that

that from the fact

for the development of

the of development for

the base of the

base of the the

for example in the

the for in example
to deal with the
deal with to the
as a matter of
a of matter as
it is important that
that important is it
the presence of the
presence of the the
the value of the
the the of value
the degree to which
to degree which the
an increase in the
in the increase an
in terms of the
terms the in of
to ensure that the
to the that ensure
is likely to be
to is be likely
are likely to be
are to likely be
the extent to which
which the to extent
the rest of the
of the the rest
the surface of the

of surface the the

at the beginning of

beginning the of at

in england and wales

england in wales and

it is not surprising

is not it surprising

in the next section

section the in next

studies have shown that

have shown studies that

an important role in

important role in an

a high proportion of

high a proportion of

are shown in figure

shown are figure in

an important part in

in an important part

has been shown that

been has that shown

is shown in fig

is in shown fig

does not mean that

that not does mean

in the united kingdom

kingdom in the united

of the nineteenth century

nineteenth century of the

will be able to

will able be to

a great deal of

great deal a of

it has been shown

shown it has been

it is unlikely that

it that unlikely is

is shown in figure

shown figure in is

as we have seen

as seen we have

in the united states

in states united the

has been suggested that

that been suggested has

has been shown to

been to shown has

on the other hand

other the on hand

it should be noted

it be should noted

on the one hand

on the hand one

a wide range of

of a wide range

should be noted that

should be that noted