8-11-2020

Developing and Testing Alternative Benchmarks of Lexical Sophistication: L2 Lexical Frequency, Semantic Context, and Word Recognition Indices

Katia Vanderbilt

DEVELOPING AND TESTING ALTERNATIVE BENCHMARKS OF LEXICAL SOPHISTICATION: L2 LEXICAL FREQUENCY, SEMANTIC CONTEXT, AND WORD RECOGNITION INDICES

by

KÁTIA MONTEIRO VANDERBILT

Under the Direction of Scott Crossley, PhD

ABSTRACT

Previous research has traditionally used first language (L1) English linguistic norms as a benchmark to assess second language (L2) production (Cook, 1992) and to select experimental stimuli in bilingual studies (Vaid & Meuter, 2017). Despite the immense contribution of this approach, L1 benchmarks may not completely represent the linguistic experience of L2 users, and they might limit our understanding of multicompetence or the state of knowing multiple languages (Cook, 1991; Klein, 1998; Vaid & Meuter, 2017). A few attempts to develop indices that more closely represent L2 linguistic experience have been made (e.g., Monteiro et al., 2020; Naismith et al., 2018), but researchers have been slow to respond to the need for more L2

benchmarks. The primary aim of this dissertation is to help address this gap by developing lexical benchmarks based on L2 corpora and L2 behavioral data collected for this dissertation. The corpus-based benchmarks included L2 lexical frequency indices, L2 range indices, and L2 semantic context indices based on Latent Semantic Analysis (LSA) and Word to Vector (Word2vec) computational methods. The benchmarks based on behavioral data included L2 word recognition indices from a word naming task performed by bilinguals studying in the United States (N = 94). These benchmarks were validated against psycholinguistic data of L2 lexical processing and human judgments of L2 writing proficiency. The results suggested that the L2 benchmarks were successful predictors of L2 writing quality and L2 word processing and were more predictive than L1 benchmarks in some cases. Analysis of individual output also suggested that the L2 benchmarks provide frequency and word recognition information that may be unique to L2 users.

INDEX WORDS: Natural Language Processing, Frequency, Semantic context, Word recognition, L2 writing, L2 benchmarks

DEVELOPING AND TESTING ALTERNATIVE BENCHMARKS OF LEXICAL SOPHISTICATION: L2 LEXICAL FREQUENCY, SEMANTIC CONTEXT, AND WORD RECOGNITION INDICES

by

KÁTIA MONTEIRO VANDERBILT

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2020

DEVELOPING AND TESTING ALTERNATIVE BENCHMARKS OF LEXICAL SOPHISTICATION: L2 LEXICAL FREQUENCY, SEMANTIC CONTEXT, AND WORD RECOGNITION INDICES

by

KÁTIA MONTEIRO VANDERBILT

Committee Chair: Scott Crossley

Committee: Eric Friginal

Kristopher Kyle

Mihai Dascălu

Ute Römer

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

August 2020

ACKNOWLEDGEMENTS

Throughout the Ph.D. program and dissertation process, I have continuously thought about the significance of privilege. I have been blessed with the privilege of having access to higher education, material resources, and, especially, people that propelled me to succeed. Without experts, friends, and family, none of my accomplishments would have been possible. I could write an endless list thanking the amazing people that motivated and inspired me from my first years of high school education until now, including brilliant minds that, for the lack of the privilege that I had, could not make as far. However, space is limited, so here are a few special words of appreciation to those who directly helped me with this dissertation.

I am thankful for Scott Crossley's expertise and guidance along the way. He pushed me to become an independent scholar, to face my deepest doubts regarding my abilities to code, and tolerated my resistance and sarcastic comments, not to mention a few tears shed in his office.

I am thankful to Mihai Dascălu and his assistant Robert Botarleanu and for their immense contribution to developing the semantic context indices tested in this dissertation. I also appreciate their assistance during the index validation phase, especially the thorough explanations provided.

I am thankful to Kristopher Kyle, who provided invaluable insights that helped improve this dissertation. I am also thankful for his generosity in sharing frequency lists and helping me with TAALES. His hard work in developing several tools has also been inspiring.

I am thankful to Ute Römer, whose contribution motivated me to look beyond the numbers and dig into the texts. Her research has also been inspiring for recognizing the expertise in L2 writing, which has shaped this dissertation in many senses.

I am thankful to Eric Friginal, whose ideas helped me to reframe the theoretical motivations of this dissertation and organize the manuscript. I am also grateful for the months we worked together and the support he has provided in the past months.

My husband, Bill Vanderbilt, also deserves many thanks. Bill has provided emotional and technical support throughout the way. He brought me back to reality when my mind was crowded with thoughts of failure, and we spent hours together troubleshooting and walking through my codes. Until today he comes to my desk in excitement, asking, “Are you working with Python?”. Finally, I could not be more thankful for his understanding when I had to close the office door and hibernate there for a while.

Lastly, I am grateful for having my parents’ support. They have taught me that hard work comes in many forms, and that this work is by no means better than any other. They also taught me that knowledge comes with the responsibility to help those who did not have the same privileges that I had.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

The lexicon is the "locus of creativity in language" (Pierrehumbert, 2012, p. 16), allowing writers, signers, and speakers to produce language sequences that have never been produced before. Perhaps, for this reason, lexical knowledge has been one of the most investigated linguistic phenomena in second language (L2) writing (e.g., Berger et al., 2017; Crossley et al., 2010, 2013; Dabbagh & Enayat, 2019; Laufer & Nation, 1995; Mazgutova & Kormos, 2015) and in L2 processing (e.g., de Groot et al., 2002; Diependaele et al., 2013; Dijkstra & Heuven, 2002; Lemhöfer & Dijkstra, 2004; Portocarrero et al., 2007). Lexical knowledge is a multifaceted construct often associated with two dimensions: breadth or the quantity of lexical knowledge, and depth or the quality of lexical knowledge (Laufer & Nation, 1995; Meara & Bell, 2001; Read, 1993)[1]. The investigation of lexical knowledge has greatly benefitted from the automatic assessment of texts through text analytics tools (e.g., Coh-Metrix, Graesser et al., 2004; VocabProfile, Heatley et al., 2002; TAALES, Kyle et al., 2018), which can account for the multifaceted nature of lexical knowledge by providing a variety of scores related to lexical complexity (e.g., density, diversity, and sophistication measures). Particularly, automatic approaches to measuring lexical knowledge afford the analysis of natural language from large corpora and the rich investigation of several lexical features concurrently, helping us understand several lexical phenomena on a scale impossible to be done manually (McNamara et al., 2017; Meurers, 2013; Meurers & Dickinson, 2017).

Several benchmarks have been developed for the automatic assessment of lexical complexity. Some of the most common benchmarks are lexical density (i.e., the proportion of

[1] See Henriksen (1999) and Qian and Schedl (2004) for alternative definitions of lexical knowledge that include other descriptors such as receptive and productive knowledge.

content words in a text; Perfetti, 1969; Read, 2000), lexical diversity (i.e., the variety of lexical items in a text; Jarvis, 2002; McCarthy & Jarvis, 2010), frequency (i.e., corpus-based rankings of lexical items based on number of occurrences; Laufer & Nation, 1995; West, 1953), range (i.e., corpus-based rankings that consider context count; Adelman et al., 2006), word information (i.e., word properties such as concreteness and familiarity as judged by humans; Coltheart, 1981), word recognition (i.e., behavioral information such as reaction time from word reading tasks; Balota et al., 2007), and contextual distinctiveness (i.e., the number of unique contexts in which a word appears; McDonald & Shillcock, 2001). These indices have been extensively used in the automatic investigation of lexical knowledge in first language (L1) and L2 writing. Overall, L2 research has suggested that more advanced users[2] produce language that is more lexically diverse (Jarvis, 2002; Yu, 2010) and with more sophisticated words that are less concrete, less familiar (Crossley et al., 2015; Crossley & McNamara, 2012; Kyle et al., 2018), less frequent (Crossley et al., 2013; Kyle & Crossley, 2015; Monteiro et al., 2020), and more difficult to process (Kyle et al., 2018). In L2 lexical processing studies, lexical sophistication has also been reported to be directly related to lexical processing. For example, words that are more frequent (Brysbaert et al., 2000; Diependaele et al., 2013), more concrete (Skalicky et al., in press), more imageable (de Groot et al., 2002), and occur in more contexts (Berger, Crossley, & Skalicky, 2019) are processed faster.

The studies mentioned above have contributed to important advancements in applied linguistics and psycholinguistics regarding L2 lexical proficiency; however, they have relied on indices derived from L1 corpora, whereas indices based on L2 corpora have not been extensively

[2] As in Cook (1992), the term L2 user was adopted instead of L2 learner or non-native speaker, and, in this dissertation, it usually refers to L2 users of English. The term L1 user was usually applied to refer to L1 users of English. This term avoids the assumption that participants in research are all actively engaged in language learning.

explored as benchmarks that represent L2 experience with English. Different from an L1, which tends to be used in several conversational contexts, an L2 can be limited to very specific purposes and particular interlocutors (Cook, 1992; Ortega, 2016; Vaid & Meuter, 2017), meaning that only domain-specific lexical knowledge may be developed. Additionally, many L2 users develop linguistic knowledge under unique circumstances where input may be limited, including limited access to native input (Ling & Braine, 2007; Ulate, 2014). These unique circumstances are particularly relevant for the study of English, which has gained the status of international lingua franca, with L2 users outnumbering L1 users (MacKenzie, 2018) and many L2 users using English to communicate exclusively with other L2 users of English (Kameda, 1992). These limitations in terms of linguistic exposure not only affect the number of lexical items L2 users of English learn (i.e., the breadth of lexical knowledge) but also the strength of these lexical representations (i.e., the depth of lexical knowledge). Therefore, L2 corpus-based indices may be needed as benchmarks that more closely represent the linguistic experience that most L2 users of English have around the globe.

Akin to corpus-based L2 indices, indices based on L2 behavioral data are scarce, with research primarily relying on L1 indices. Psycholinguistic research has repeatedly reported important quantitative differences between monolingual and bilingual[3] processing (Bialystok, 2009). Overall, studies have found a deficit in retrieval and processing among bilinguals, as evidenced by research showing a response lag in word reading tasks (de Groot et al., 2002; Diependaele et al., 2013; Monaghan et al., 2017), more tip of the tongue issues (Gollan & Acenas, 2004), and a smaller vocabulary size compared to monolinguals (Portocarrero et al.,

[3] The terms bilingual and L2 users are used in this dissertation interchangeably to mean a user (i.e., speaker, writer, signer) of a language other than their first language. For simplicity purposes, these terms also refer to users of English as a third, fourth, or fifth language (i.e., multilinguals).

2007). These disadvantages are often resolved when frequency is accounted for, suggesting that it is the reduced experience with lexical items that cause processing delays (Bialystok, 2009; Johns et al., 2016). It is worth noting that qualitative differences have not been found between monolinguals and bilinguals. For example, neuroscience and psycholinguistic studies have suggested that both languages are processed in the same region of the brain (see reviews by Perani & Abutalebi, 2005; Steinhauer, 2014) and that the L1 and L2 lexicons operate similarly and conjointly (Brysbaert et al., 2000; Monaghan et al., 2017). Notwithstanding the similarities, when both languages are considered, a bilingual will never match the performance of two monolinguals (Bialystok, 2009). Therefore, L2 word recognition indices may be used as alternative benchmarks that represent L2 processing.

One solution to address the lack of indices that provide a closer representation of L2 processing and linguistic experience is to broaden the automatic lexical benchmarks available to sample L2 corpora and L2 processing data. This has been done on a smaller scale in previous studies that have developed corpus-based L2 indices (Monteiro et al., 2020; Naismith et al., 2018). This dissertation is a step towards expanding this research agenda by adding new L2 automatic indices collected on larger scales that more directly represent L2 experiences. Specifically, four types of automatic indices were developed from L2 corpora and L2 behavioral data: lexical frequency, range, semantic context, and word recognition information. The frequency, range, and semantic context indices were developed from the L2 written corpus EF-CAMDAT (English First-CAMbridge Open Language Database; Huang et al., 2017), which can be used to indirectly represent the linguistic experience of English foreign language learners across the globe. Specifically, EF-CAMDAT provides an indirect representation of the experience of learning through writing in an online classroom environment. EF-CAMDAT was

selected for being one of the largest L2 corpora available and for representing both the production of L2 users and the language to which they were exposed through classroom tasks.[4] The word recognition indices were based on L2 behavioral data collected for this dissertation from a word naming task (i.e., a word reading psycholinguistic task) from bilinguals studying in the United States, most of whom had limited experience with English. While recognizing that the bilingual experience is too broad to be contained in one corpus or one psycholinguistic experiment, the indices developed for this dissertation can certainly contribute to the expansion of indices that represent L2 experiences with English.

The primary aim of this dissertation is to test the validity of these L2 indices as benchmarks of lexical sophistication through a series of models that test the predictive power of the indices by themselves and in the presence of similar indices. Developing and validating indices that represent different experiences with the input has been one of the major challenges in lexical proficiency research, and an endeavor that has contributed immensely to the advancement of lexical proficiency research (Adelman et al., 2006; Heuven et al., 2014; Mandera et al., 2017). In the well-cited article by Heuven et al. (2014), for example, indices based on subtitles of television programs (i.e., SUBTLEX-UK) were found to be stronger predictors of lexical decision data than indices based on the British National Corpus, composed of written and spoken samples from a variety of sources. The authors claimed that subtitles may be used as a proxy of spoken language that may be more representative of the linguistic experience of many language users. Many studies have followed suit and successfully used

[4] It is worth pointing out that EF-CAMDAT incorporates not only the language that participants naturally produced, but also language from classroom tasks. Because the indices represent the indirect linguistic experience of L2 users, the inclusion of task input is not problematic given that task input also represents the L2 experience with language. Also, it would be impossible to gauge whether the lexical items from the tasks were spontaneously produced by the learners as a result of learning from the tasks or a direct copy of the input.

indices based on subtitles as predictors of language production and processing (e.g., Berger et al., 2019; Crossley & Salsbury, 2010; Mainz et al., 2017). This dissertation follows this important research tradition of testing the validity of new benchmarks as measures of lexical proficiency by investigating the power of the L2 indices as explanatory variables of L2 writing and L2 lexical processing. For direct comparisons, similar L1 indices were tested, and the explanatory power of the indices was compared. Like previous research, comparisons with similar indices allow us to test whether new benchmarks can provide explanatory power beyond what other available indices can offer (Heuven et al., 2014; Mandera et al., 2017). Due to the limited availability of robust and comparable L2 indices, a comparison with L1 indices was judged more appropriate.

The validation of the indices was performed in two major steps. In the first validation step, the lexical frequency, context diversity, semantic context, and word recognition indices were used as explanatory variables of writing proficiency data from the integrated (N = 480) and independent task (N = 480) from the TOEFL iBT. This type of validation has been extensively used in L2 writing studies (e.g., Crossley et al., 2010; Crossley & McNamara, 2012; Monteiro et al., 2020). The explanatory power of these variables was tested against the predictive power of similar L1 frequency and range variables as available in the Tool for the Automatic Analysis of Lexical Sophistication (TAALES, Kyle et al., 2018; Kyle & Crossley, 2015). This procedure, which has also been adopted in recent studies that have used L2 norms (Monteiro et al., 2020; Naismith et al., 2018), was meant to test whether the L2 norms had predictive power beyond similar L1 norms. In the second validation step, the L2 lexical frequency, range, and semantic context indices were used to predict the L2 behavioral data (i.e., L2 lexical decision task data) publicly released by Berger, Crossley, and Skalicky (2019). This procedure is similar to psycholinguistic studies such as Diependaele et al. (2013), Brysbaert et al. (2017), and Johns et

al. (2016), which have used similar L1 norms to the development of L2 lexical processing models. The explanatory power of these variables was also tested against the explanatory power of similar L1 word recognition variables as available in TAALES. These validation steps address the three overarching questions of this dissertation:

1. To what degree are L2 and L1 lexical frequency and range indices derived from written corpora predictive of L2 writing scores and L2 lexical processing data?
2. To what degree are L2 and L1 semantic context indices derived from written corpora predictive of L2 writing scores and L2 lexical processing data?
3. To what degree are L2 and L1 word recognition indices derived from behavioral data comparable and predictive of L2 writing scores?

This dissertation is organized in three main studies, hereafter referred to as Study 1, Study 2, and Study 3. These studies are reported in chapters 2, 3, and 4, respectively. Each of these studies addresses each research question and contains the traditional parts of a research article: literature review, methods, results, and discussion. Chapter 5 includes a summary of all results, discusses differences between the L2 and L1 indices, and considers a few implications for the fields of applied linguistics and psycholinguistics. The chapters are outlined below.

Chapter 2 contains Study 1, which answers the first question of this dissertation regarding the predictive power of the L2 frequency and range indices. The literature review includes studies on the impact of frequency and range in L2 writing and L2 lexical processing. The methods section describes how the L2 frequency and range indices were developed and the L1 indices used for comparison purposes. It also includes a description of the TOEFL data used to build the L2 writing models in studies 1, 2, and 3 and the L2 lexical decision data from Berger, Crossley, and Skalicky (2019) used to build the L2 lexical processing models in Study 1 and 2.

The results section reports on the statistical models testing the predictive power of the L2 and L1 indices. The statistical models are divided into two main parts: the L2 writing models and the L2 lexical processing models. The discussion section explains the findings and discusses implications.

Chapter 3 contains Study 2, which answers the second research question regarding the predictive power of the L2 semantic context indices. The literature review includes studies that have used semantic context information to the analysis of L2 writing and L2 lexical processing. The methods section describes the computational methods used to develop the L2 and L1 indices (i.e., LSA and Word2vec) and each L2 and L1 index included in this dissertation. The results section reports on the L2 writing and L2 lexical processing statistical models. The discussion section explains the findings and discusses implications.

Chapter 4 contains Study 3, which answers the third research question regarding the validity and predictive power of the L2 word recognition indices. The literature review explores L2 word processing studies and L2 writing studies that have used behavioral-based indices as benchmarks. The methods section describes the word recognition task used in this dissertation (i.e., word naming task) and the L2 indices that were built from this dataset. The results section reports on comparisons between the word naming data and similar datasets publicly available and the L2 writing models. The discussion section addresses the findings and implications.

Chapter 5 concludes this dissertation by providing a comparison of the findings across the studies, a comparison of the L1 and L2 indices, and a discussion of the implications and possible applications of the indices in applied linguistics and psycholinguistics. Future directions for the incorporation of L2 indices in Natural Language Processing (NLP) are also provided.

2 STUDY 1: DEVELOPING AND TESTING L2 LEXICAL FREQUENCY INDICES

Second language users have to acquire thousands of words and multi-word expressions to become fluent speakers and writers, making lexical proficiency a major component of language learning (Laufer & Shmueli, 1997). Lexical proficiency, as a linguistic and cognitive phenomenon, has been particularly important in studies of L2 writing quality (Biber & Gray, 2013; Friginal et al., 2014; Kyle & Crossley, 2016; Römer, 2009b) and models of bilingual processing (Brysbaert et al., 2017; Dijkstra & Heuven, 2002; Skalicky et al., in press). These two research areas have made important contributions to our understanding of lexical proficiency. The research on L2 writing has contributed to our understanding of how lexis develops over time (e.g., Crossley et al., 2019), the similarities between L2 and L1 writing (Römer, 2009b), and the relationship between L2 writing quality assessment and lexical sophistication (e.g., Kyle & Crossley, 2016), among many others. The research on bilingual processing has answered important questions regarding the integratedness of the bilingual lexicon, the existence of L1 interference on the L2 lexicon, and the degree of influence of L2 proficiency on L2 lexical processing (Balota et al., 2007; Dijkstra & Heuven, 2002; Vanlangendonck et al., 2019).

Previous studies, such as the ones described above, have relied on lexical complexity benchmarks or indices that gauge several dimensions of lexical proficiency to investigate L2 writing and L2 lexical processing. Among these indices are lexical density (Laufer & Nation, 1995), lexical diversity (Laufer & Nation, 1995), lexical bundles (i.e., frequent word combinations; Cortes, 2004), phrase frames (i.e., productive patterns with a variable slot; O'Donnell et al., 2013), and lexical frequency (Ellis, 2002). Many of these indices rely on reference corpora, which are taken to represent the naturally occurring language to which a group of speakers or writers is exposed. Due to the limited availability of L2 corpora, these

indices have been predominantly based on corpora primarily from L1 production, with a few exceptions (Monteiro et al., 2020; Naismith et al., 2018). With the increasing availability of large sets of L2 corpora, the development of L2 automatic indices that represent the language produced by L2 users, to which many foreign L2 speakers are exposed (Ling & Braine, 2007; Ulate, 2014), is possible. Automatic lexical indices derived from L2 corpora may help us understand L2 production beyond that offered by L1 indices (Crossley et al., 2019; Naismith et al., 2018) and afford the opportunity to replicate analyses with different corpora to test the strength of past conclusions based on L1 benchmarks (Bestgen, 2017; Porte, 2012). Study 1 of this dissertation sets out to contribute to the development of automatic indices that represent L2 language by developing two types of L2 frequency indices: lexical frequency and range. Both indices for single lemmas and n-gram lemmas (i.e., bigrams and trigrams) are developed and validated in this study.

The validation of the L2 indices included the replication of psycholinguistic and applied linguistic studies that have used L1 lexical sophistication benchmarks as explanatory variables of L2 production. Specifically, the L2 automatic indices were used to generate a lexical profile of L2 texts to model L2 writing proficiency. A second validation step included the automatic analysis of words from a behavioral task to model L2 lexical processing. These validations are meant to answer the first research question of this dissertation regarding the usefulness of the L2 indices as predictors of L2 writing and L2 word processing.

2.1 Lexical Sophistication

Lexical sophistication is a component of lexical complexity, and it is often associated with the depth and breadth of lexical knowledge (Laufer & Nation, 1995; Meara & Bell, 2001). Several indices have been proposed to the investigation of lexical sophistication, including

indices that measure psycholinguistic properties such as word concreteness (Coltheart, 1981) and word properties such as length and orthographic neighbors (Balota et al., 2007). Perhaps the most common measure of lexical sophistication is corpus-driven frequency (Ellis, 2002), which provides rankings for lexical items in reference to a representative corpus. A derivative measure of frequency recently featured in the literature is range, also referred as contextual diversity (Adelman et al., 2006), which represents the frequency of texts in which lexical items appear. These indices, which have been primarily based on corpora containing L1 texts, have been used extensively to generate a lexical profile of essays and words which are then used as explanatory variables of L2 writing quality and L2 lexical access (e.g., Biber & Gray, 2013; Brysbaert et al., 2017; Cumming et al., 2006; Guo et al., 2013; Johns et al., 2016; Kyle & Crossley, 2016). The literature below details the role of frequency and range on the L2 writing and L2 lexical processing literature, with a focus on studies that have used automatic lexical indices.

2.2 Lexical Frequency

Psycholinguistics and cognitive linguistics propose that a major driving force in language acquisition is the repeated exposure to linguistic forms (Ellis, 2002; Ellis et al., 2016). Ellis (2002) argues that the human brain is tuned to the frequency of lexical and lexical-grammatical features, being able to abstract regularities such as phrase frames (e.g., *it is** + adj) and grammatical rules. Lexical access is also facilitated by frequency in the input (Brysbaert et al., 2000; Diependaele et al., 2013), with more frequent words being retrieved and produced faster. The repeated exposure afforded by frequency in the input strengthens the mental representations of lexical items, allowing for more efficient processing. Despite the undeniable influence of repeated exposure, the so-called “frequency effect” has been criticized for being better fitted to native language acquisition, for which it was originally envisioned (Gass & Mackey, 2002). Gass

and Mackey argue that frequency is only one factor influencing L2 linguistic development and that other factors related to perceptual salience, semantic complexity, morphological regularity, explicit instruction, awareness, and L1-transfer are as relevant. However, despite some evidence that the frequency effect may not be pronounced in earlier stages of language learning (Crossley et al., 2010; Crossley, Skalicky, et al., 2019; González, 2017), there is plenty of evidence suggesting the existence of a frequency effect both in L2 writing (e.g., Crossley & McNamara, 2012; Johnson et al., 2016; Kyle & Crossley, 2015; Laufer & Nation, 1995; Meara & Bell, 2001; Römer, 2016) and in L2 word processing (Brysbaert et al., 2000, 2017; Diependaele et al., 2013; Lemhöfer et al., 2008). It then seems that, when intervening variables are taken into consideration, frequency is a fundamental cognitive mechanism that permeates all domains of linguistic processing, including among L2 users. The influence of lexical frequency on L2 writing and L2 lexical processing studies is reported below.

2.2.1 Lexical Frequency and L2 Writing

Lexical frequency has been featured in several studies of L2 writing, especially in studies that have used automatic lexical indices. In these studies, lexical frequency is used as a proxy of lexical sophistication, with more frequency related to less sophistication. L2 writing studies have suggested that words with higher frequency are more common in lower-level writing (Crossley et al., 2015; Guo et al., 2013; Kyle & Crossley, 2015; Palfreyman & Karaki, 2019). By the same token, proficient learners use words with lower frequency, which are considered more sophisticated. However, there is also evidence that lower-level writing can contain a high incidence of low-frequency words (e.g., Crossley et al., 2010, 2019; González, 2017), a fact partially attributed to possible frequency effects from the L1. Regarding the use of multi-word combinations, the results are somewhat contradictory. There is both evidence that less

experienced L2 writers use few frequent word combinations or lexical bundles from L1 writing (e.g., Ädel & Erman, 2012; Shin et al., 2018) and others suggesting more use of lexical bundles in lower levels (e.g., Bychkovska & Lee, 2017; Staples et al., 2013). Research also suggests that more experienced L2 writers such as senior undergraduate students and graduate students produce n-grams in a similar fashion as L1 writers (e.g., O'Donnell et al., 2013; Römer, 2009b). In statistical models of writing quality, the effect of n-gram frequency on writing quality is usually positive (e.g., Kyle & Crossley, 2016, 2015; Monteiro et al., 2020), meaning that more proficient writers use more common word combinations, which are possibly more idiomatic.

2.2.2 Lexical Frequency and L2 Lexical Processing

Important questions regarding L2 lexical processing have been answered by word recognition tasks such as lexical decision, a task requiring participants to judge whether a string of letters is a word or non-word, and word reading tasks (Balota et al., 2007; Dijkstra & Heuven, 2002; Vanlangendonck et al., 2019). Because single words can be easily manipulated in experiments, they are useful for investigating processing phenomena (Balota et al., 2007). This research has primarily focused on L1 users (e.g., Balota et al., 2007; Yap et al., 2012), with increasing research on L2 users (Berger, Crossley, & Skalicky, 2019; Brysbaert et al., 2017; Fender, 2003; Lemhöfer et al., 2008), and has provided insights into the characteristics of words that influence this processing through statistical models.

Models of word processing are developed by using word characteristics such as cognate status (Vanlangendonck et al., 2019), word length (Berger, Crossley, & Skalicky, 2019; de Groot et al., 2002; Skalicky et al., in press), and lexical frequency (Diependaele et al., 2013; Duyck et al., 2008) as explanatory variables of word processing behavior (e.g., reaction time to a stimulus word). Word frequency has been featured in several of these studies (e.g., Brysbaert et al., 2017;

Diependaele et al., 2013; Duyck et al., 2008; Lemhöfer et al., 2008; Van Wijnendaele & Brysbaert, 2002), which have found that words that are less experienced by L2 users, as measured by frequency indices, are less entrenched in their mental lexicon. Word processing studies have also found that frequency has a stronger effect in L2 processing than in L1 processing (e.g., Brysbaert et al., 2017), and that higher proficiency in the L2 results in weaker frequency effects (Brysbaert et al., 2017; Lemhöfer et al., 2008). These findings have important implications for the understanding of the bilingual lexicon. The frequency effect, for example, suggests that the delay in processing among bilinguals is less related to a lack of neural plasticity in the adult bilingual brain than with limited linguistic experience (Morrison et al., 2002).

2.3 Range

Despite extensive evidence on the effect of repeated exposure in lexical proficiency, frequency has been criticized for being confounded with other variables such as range (Adelman et al., 2006). Some scholars argue that lexical development is primarily affected by repeated encounters spaced across contexts (Adelman et al., 2006; Baayen, 2010; Verkoeijen et al., 2004). The argument is that spaced repetitions in multiple contexts can have a facilitating effect on memorization because the lexical items are accessed or activated more frequently, strengthening their mental representations while also strengthening the connections with related lexical items in the mental lexicon (Adelman et al., 2006; Glenberg, 1979). Therefore, indices that represent the number of contexts or texts in which lexical items appear may better represent the linguistic experience with input than absolute frequency counts (Adelman et al., 2006). Although not as extensively investigated as an index of linguistic experience and lexical sophistication, measures of range have been shown to be promising predictors of L2 writing and L2 lexical processing, as detailed below.

2.3.1 Range and L2 Writing

Different from frequency indices, which have been extensively tested as benchmarks of lexical sophistication, range indices have been featured in the L2 writing literature only recently. These recent studies have suggested that higher-level learners use words that appear in fewer texts (Kyle et al., 2018; Kyle & Crossley, 2016, 2015; Monteiro et al., 2020), regardless of whether the indices were based on L1 corpora (Kyle et al., 2018; Kyle & Crossley, 2016, 2015) or L2 corpora (Monteiro et al., 2020). The trend seems to be the opposite for n-grams, with studies finding that higher-level learners rely on word combinations that occur in more texts (Garner et al., 2019; Monteiro et al., 2020). Monteiro et al. (2020) argue that the use of common n-grams, even when less sophisticated, may signal idiomatic knowledge, which is perceived positively by raters of L2 writing.

2.3.2 Range and L2 Lexical Processing

Range measures have also been successfully used to explain L2 lexical processing behavior (Berger, Crossley, & Skalicky, 2019; Hamrick & Pandža, 2020; Johns et al., 2012; Skalicky et al., in press). Some of these studies have found a clear advantage for range over frequency in L2 processing (Johns et al., 2012; Skalicky et al., in press), while others have found an effect for both frequency and range (Hamrick & Pandža, 2020). Jones et al. (2017) explain that it is the syntactic and morphological co-occurrence probabilities of words that cause frequency measures to be strong predictors of lexical processing. The authors add that while frequency is based on the "principle of repetition," a classic but fragmentary principle of learning and memory, range is based on the "principle of likely need" which establishes that words that are present in more contexts are likely needed in others, being accessed more frequently and developing stronger connections with other related words. Jones et al. (2017) conclude that

ultimately it is the distributional properties of words that assist with lexical development, organizing the mental lexicon.

2.4 Research Questions

Study 1 addresses the first research question of this dissertation regarding the predictive power of the L2 lexical frequency and range indices derived from written corpora as explanatory variables of L2 writing scores and L2 lexical processing data by themselves and in comparison with similar L1 indices. The following specific research questions guided Study 1:

1) To what extent are L2 and L1 lexical frequency and range indices derived from written corpora predictive of L2 writing proficiency?
2) To what extent are L2 and L1 lexical frequency and range indices derived from written corpora predictive of L2 lexical decision reaction time and accuracy scores?

2.5 Methods

Study 1 uses frequency and range L2 indices as predictor variables of L2 writing proficiency and L2 behavioral data (i.e., reaction time and accuracy values) from a lexical decision task. The predictive power of the L2 frequency and range indices was compared to the predictive power of similar L1 indices. These indices included frequency (i.e., number of lemmas and lemma n-grams in the corpus) and range values (i.e., number of texts in which each lemma and lemma n-grams occurred) developed from an L2 corpus. The L2 corpus, independent variables (i.e., L2 and L1 frequency and range norms), and dependent variables (i.e., test scores from the TOEFL iBT, and reaction time and accuracy scores) are outlined below.

2.5.1 EF-CAMDAT Indices

The L2 frequency and range indices were derived from the English First-CAMbridge open language database (EF-CAMDAT; Huang et al., 2017). Indices for all lemmas, content

lemmas, function lemmas, lemma bigrams, and lemma trigrams were developed. The corpus and the indices are detailed below.

2.5.1.1 EF-CAMDAT Corpus

The EF-CAMDAT (Huang et al., 2017) is a large corpus of written data produced by 174,743 L2 users from 198 nationalities at multiple proficiency levels (see Table 2.1 below). It includes written samples from a variety of learner writing tasks from the online English course *Englishtown*. A beginner level task, for example, requires students to write an e-mail introducing themselves, and an advanced task requires students to retell a news story. This corpus was selected for being the largest L2 corpus available, for representing a common type of linguistic experience (i.e., classroom-based writing), and for including a variety of topics and tasks, potentially resembling linguistic exposure in real world tasks. While the EF-CAMDAT has essays from varying levels, levels B and C, as defined by the Common European Framework of Reference for languages (CEFR), were used. The texts from A1 and A2 levels were short and contained many misspellings and ill-formed sentences, making them unsuitable for text processing. Table 2.1 below, reproduced from Huang et al. (2017), illustrates the *Englishtown* levels in relation to standardized tests such as TOEFL and IELTS. Levels B and C, which encompasses levels 7 to 16 in *Englishtown*, contained a total of 30,771,991 words from 246,328 texts. Table 2.2 shows the number of words and texts per the *Englishtown* level.

Table 2.1 Englishtown Skill Levels in Relation to Common Standards from Huang et al. (2017)

Englishtown	***1–3***	***4–6***	***7–9***	***10–12***	***13–15***	***16***
Cambridge ESOL	–	KET	PET	FCE	CAE	–
IELTS	–	<3	4–5	5–6	6–7	>7
TOEFL iBT	–	–	57–86	87–109	110–120	–
TOEIC Listening & Reading	120–220	225–545	550–780	785–940	945	–
TOEIC Speaking & Writing	40–70	8–110	120–140	150–190	200	–
CEFR	A1	A2	B1	B2	C1	C2

Table 2.2 EF-CAMDAT Number of Words and Texts by Level

Level	***Number of Texts***	***Number of Words***	***Number of Words per Text***
Level 16	1,940	375,664	193.64
Level 15	2,236	427,016	190.97
Level 14	3,631	695,658	191.59
Level 13	8,831	1,646,674	186.46
Level 12	9,256	1,598,429	172.69
Level 11	15,588	2,569,312	164.83
Level 10	36,485	5,107,376	139.99
Level 9	28,553	3,461,275	121.22
Level 8	41,926	4,707,024	112.27
Level 7	97,882	10,183,563	104.04
Total	246,328	30,771,991	124.92

2.5.1.2 L2 Frequency and Range Indices

Two types of lexical frequency indices were developed using the EF-CAMDAT corpus: lexical frequency and range. Frequency was operationalized as the rate at which a given lexical item appears in a representative corpus and range as the frequency of texts in which a given lexical item appears. For example, the word “able” appeared 1,567 times (i.e., frequency) in 1,328 texts (i.e., range). To create the frequency and range lists from the L2 corpora, the programing language *Python* (van Rossum, 1995), along with the *Pandas* libraries (McKinney et al., 2010), and *NLTK* suite of libraries (Loper & Bird, 2002) were used.

Five types of frequency and range indices, representing multiple types of lexical representations, were developed. All these indices were developed from lemmas, as opposed to raw frequencies. Lemmas are inflected forms of the same base (e.g., study from study, studied, studying, studies). These include indices for content lemmas, function lemmas, all lemmas, bigram lemmas, and trigram lemmas. Frequencies are represented as lemmas to account for theories that words are stored as lemmas (Jiang, 2000). Besides, lemmatization allows inflectional variants to be collapsed, increasing the distributional information to be added to

statistical models (Riordan & Jones, 2011). All words were lemmatized using Someya's (2008) lemma list. Each index type developed for this dissertation is detailed below.

2.5.1.2.1 All Lemmas

Frequency and range indices with all lemmas (i.e., content and function lemmas) were developed. Word frequency and range indices have been used in the literature for indirectly representing the language that L2 users experience (Ellis, 2002) as well as a proxy of lexical sophistication (e.g., Kyle & Crossley, 2016). Research indicates they are strong predictors of L2 development (e.g., Crossley et al., 2014) and L2 lexical processing (e.g., Diependaele et al., 2013).

2.5.1.2.2 Content Lemmas

Content words are lexical items that carry most of the meaning in utterances. They are nouns, adjectives, adverbs, and most verbs. Content lemmas were derived from the EF-CAMDAT using the *NLTK* library by eliminating the function lemmas from the corpus. Content words have been extensively featured as significant predictors of L2 writing (e.g., Crossley & McNamara, 2012; Guo et al., 2013).

2.5.1.2.3 Function Lemmas

Determiners, auxiliaries, prepositions, conjunctions, auxiliary verbs, and pronouns are function words which predominantly indicate meaning relationships (Biber et al., 2002). Function lemmas were computed by using the stopwords list from *NLTK*. Although they do not represent stylistic processes to the degree content words do, they are featured in text analysis because they may measure the successful use of referential language. Function words have been found to have a weak but significant relationship with writing quality (Kyle & Crossley, 2016, 2015)

2.5.1.2.4 Lemma N-grams

N-grams are combinations of words. Two-lemma combinations (i.e., bigrams) and three-lemma combinations (i.e., trigrams) frequency and range lists were calculated. Research has indicated that frequent n-grams may be stored as single units in the mental lexicon (Hoey, 2005). Also, n-gram frequency can explain human scores of lexical proficiency (Kyle et al., 2018; Kyle & Crossley, 2015). Lemma n-grams were calculated using the count vectorizer from *NLTK*.

2.5.1.3 Token Selection and Transformations

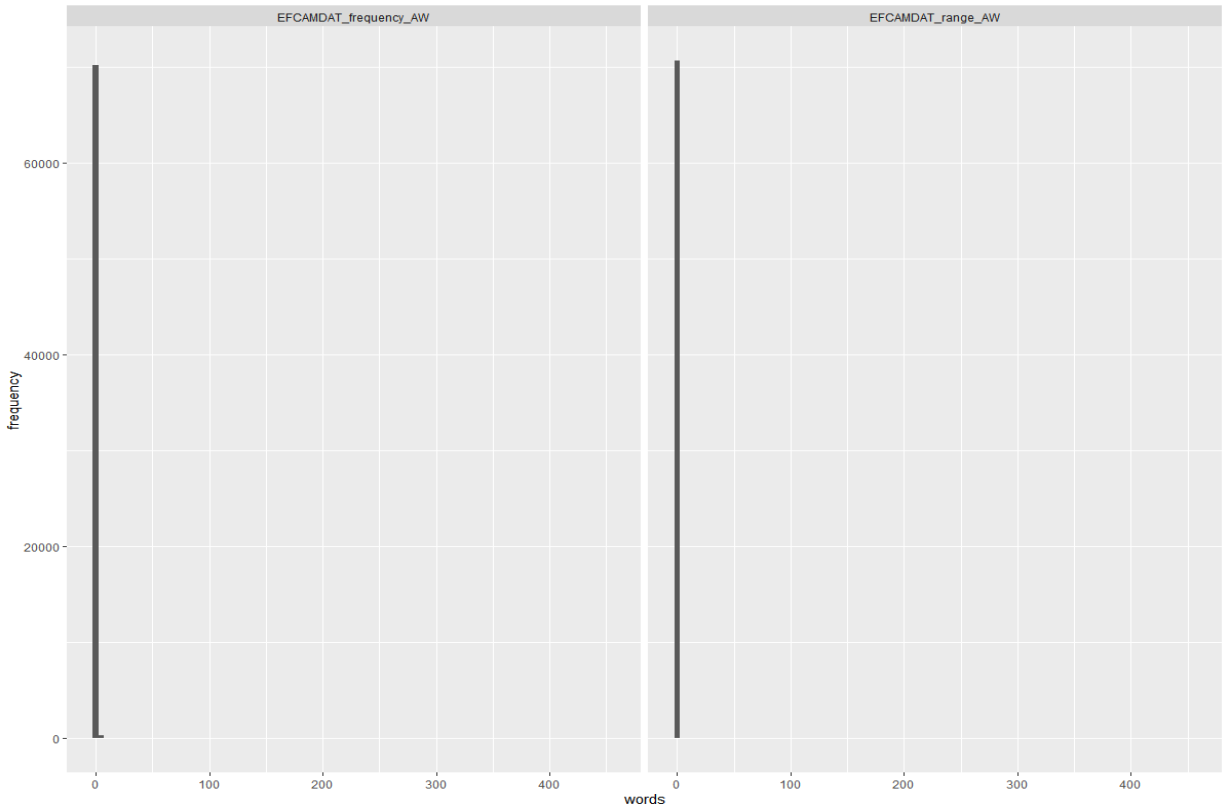
After the frequency lists were calculated, all tokens with a raw frequency of one in the EF-CAMDAT corpus were removed to reduce rare misspellings. This decision was based both on the literature and on model comparisons with lists with a more conservative cut-off point[5]. Although researchers disagree on the minimum cut-off point to use (Baron et al., 2009), Scott (1997) suggested a threshold of two in studies where corpora are compared. Because this dissertation compares indices from different corpora, Scott's suggestion was adopted. Also, as revealed by a qualitative analysis of individual output, the less conservative cut-off point of two provided more frequency information about on-target lemmas and n-grams than off-target lemmas and n-grams (see discussion and appendices for examples). The data were normalized and subsequently log transformed. To normalize the data, the frequency of words and texts was divided by the number of words in the corpus and multiplied by 10,000.

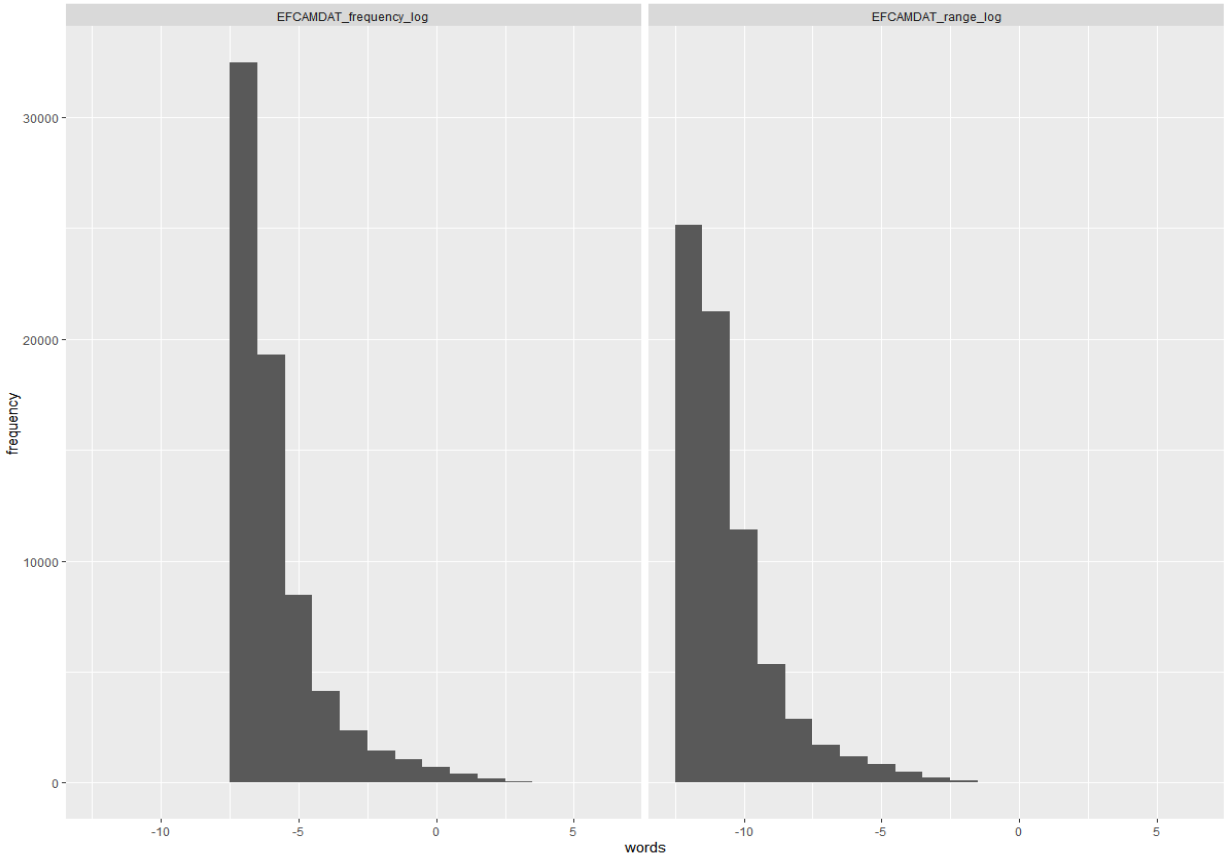
Logarithmic transformations were performed to reduce the skewness of all range and frequency indices, which were all near-Zipfian, using the natural logarithm function from *NumPy*

[5] L2 writing models that used the EF-CAMDAT frequency and range lists with a more conservative cut-off point of five (i.e., only words with a raw frequency of five or more were included) were run and compared with the models that used the lists with a cut-off point set at two. The integrated model with a more conservative cut-off point explained an additional 0.9% of the integrated scores, but the independent model with a more conservative cut-off point explained 0.7% less of the independent scores. No statistical differences between the more conservative and less conservative models were found.

(Oliphant, 2006). This function calculates the inverse of the exponential function (log(exp(x)) = x). Because transformations were performed on the normalized lists, which contained values below 1, negative log-transformed scores were generated (see discussion and individual output below). The transformation is illustrated in Figure 2.1, which shows the histograms before and after the logarithmic transformation of the EF-CAMDAT frequency for all lemmas and the EF-CAMDAT range for all lemma indices.

Figure 2.1 Histograms for the EF-CAMDAT Frequency (Left) and Range (Right) Indices for All Lemmas before and after Logarithmic Transformation





2.5.2 TAALES Indices

To test the validity of the EF-CAMDAT indices, they were compared with similar L1 indices computed by the Tool for the Automatic Assessment of Lexical Sophistication (TAALES; Kyle et al., 2018). The tool includes frequency and range information for all lemmas, content lemmas, function lemmas, bigram lemmas, and trigram lemmas similar to the ones developed for this dissertation. The COCA (Corpus of Contemporary American English; Davies, 2008) Fiction frequency and range indices from TAALES were used as the L1 benchmarks to judge the L2 indices. Models with COCA Academic were also tested, but they were, overall, less powerful than both the COCA Fiction and the EF-CAMDAT models. Due to space constraints,

the COCA Academic models are not reported.[6] COCA Fiction is detailed below.

2.5.2.1 COCA Fiction

The fiction section of COCA is composed of texts from literary magazines, popular magazines, children's books, movie scripts, first chapters of first edition books, and fan fiction (Davies, 2009). Therefore, COCA Fiction represents a range of reading experiences for a range of age groups, including children and teenagers. Also, indices based on COCA Fiction have been found to be the strongest predictors of word processing behavior when compared to other COCA registers (Brysbaert et al., 2012) and one of the strongest predictors of word choice scores in narrative essays (Kyle et al., 2018). For representing a variety of reading experiences, this corpus was judged a fair candidate for comparisons with EF-CAMDAT, which offers "a variety of receptive and productive tasks" (Huang et al., 2017, p. 3). The indices from TAALES were based on texts from 1990 to 2015 from COCA Fiction (Kyle et al., 2018). A cut-off point of 5 was adopted for the development of the COCA Fiction indices.

2.5.3 Summary of Indices

Table 2.3 below summarizes the frequency and range indices developed for this dissertation (hereafter called EF-CAMDAT indices), along with the frequency and range indices from TAALES used for statistical comparisons. Correlations between the EF-CAMDAT and COCA Fiction ranged from .79 for function words to .38 for trigrams (see Appendix A for all correlation coefficients). Correlation for content words was .64, suggesting that the corpora were similar, but that there were also differences. To illustrate, an analysis of the top 1000 words, which usually account for 80–85% of TOEFL essay words (Biber & Gray, 2013), overlapped by

[6] Models with COCA Magazine were also developed in post-hoc analyses. The COCA Magazine models explained 14% of the independent scores, and 2% of the integrated scores. Similar to the COCA Fiction models, single lemma indices were stronger predictors than n-gram-based indices. The differences between the COCA Magazine and EF-CAMDAT models were also not statistical.

548 words. The most frequent words in the top 1,000 EF-CAMDAT list, which were not present in the top 1,000 COCA Fiction list, included “experience,” “bowling,” “song,” “hi,” “study,” and “improve.”

Table 2.3 EF-CAMDAT and COCA Fiction Frequency and Range Indices

Category	***Indices***	***COCA Fiction (TAALES)***
Lexical frequency	EF-CAMDAT Frequency – All Lemmas	COCA Fiction Frequency – All Lemmas
	EF-CAMDAT Frequency – Content Lemmas	COCA Fiction Frequency – Content Lemmas
	EF-CAMDAT Frequency – Function Lemmas	COCA Fiction Frequency – Function Lemmas
	EF-CAMDAT Frequency – Lemma Bigrams	COCA Fiction Frequency – Lemma Bigrams
	EF-CAMDAT Frequency – Lemma Trigrams	COCA Fiction Frequency – Lemma Trigrams
Range Indices	EF-CAMDAT Range – All Lemmas	COCA Fiction Range – All Lemmas
	EF-CAMDAT Range – Content Lemmas	COCA Fiction Range – Content Lemmas
	EF-CAMDAT Range – Function Lemmas	COCA Fiction Range – Function Lemmas
	EF-CAMDAT Range – Lemma Bigrams	COCA Fiction Range – Lemma Bigrams
	EF-CAMDAT Range – Lemma Trigrams	COCA Fiction Range – Lemma Trigrams

2.5.4 Outcome Variables

The validation of the EF-CAMDAT indices occurred in two steps. In the first step, the indices were used as explanatory variables of writing quality. Specifically, they were tested as predictors of TOEFL iBT scores from the integrated and independent writing tasks. In the second step, the indices were used as explanatory variables of lexical processing. Specifically, the frequency and range indices were used as predictors of accuracy and reaction time from a lexical decision task from Berger, Crossley, and Skalicky (2019). The outcome variables are detailed below.

2.5.4.1 TOEFL Essays

TOEFL essays from the TOEFL iBT public use dataset were utilized. This dataset includes essays and their scores from the independent (N = 480) and integrated task (N = 480). These scores were assigned by expert raters trained by ETS who followed a holistic rubric that ranged from 0 to 5 points. Inter-rater reliability of r = .65 (Enright & Quinlan, 2010) and r = .77 (Zhang, 2008) has been reported. The rubric for each task, which can be found at https://www.ets.org/s/toefl/pdf/toefl_writing_rubrics.pdf, was based on investigations of raters' cognitive processes (Brown et al., 2005; Cumming et al., 2006), and the tasks were shown to reflect college-level writing (Biber & Gray, 2013; Riazi, 2016).

2.5.4.1.1 Independent Task

The independent task entails impromptu writing on a selected topic under time constraints (i.e., 30 minutes). This task has been used to gauge L2 users' academic writing ability by requiring test-takers to provide argumentation based on their prior knowledge of the topic. The TOEFL iBT public use dataset includes two topics: career choice (N = 240) and cooperation (N = 240). The first topic required favorable or critical arguments regarding a career choice based on a field of study or personal interest. The second topic required arguments related to the importance of cooperation in today's world compared to the past. A minimum of 300 words is recommended in this task. A high score in the independent task is assigned to an essay that is on-topic, well-organized, well-developed, coherent, and unified. At the language level, raters expect syntactic variety, appropriate word choice, and proper use of idioms.

2.5.4.1.2 Integrated Task

The integrated task consists of a written response to source texts in written and oral format under time constraints (i.e., 20 minutes). It tests the ability to select, organize, and

synthesize relevant information. The TOEFL iBT public use dataset includes two topics for this task: bird migration (N = 240) and fish farming (N = 240). The first topic required participants to summarize and critique different theories about how birds orient themselves when migrating. The second topic required a summary and a contrast of views on the effects of fish farming. The writers were allowed to take notes and were recommended to write 150–225 words. A high score in the integrated task is assigned to a response that contains key information from the sources with coherence and accuracy. Organization and little language error are also expected.

The integrated and independent scores had a moderate to strong correlation (r = .69). Table 2.4 shows the distribution of participants per score, showing that the corpus is representative of a range of writing proficiency levels.

Table 2.4 Distribution of Participants per Score for the Independent and Integrated Task

Score	***Independent Task***	***Integrated Task***
1	3	47
1.5	5	32
2	38	49
2.5	57	49
3	120	70
3.5	85	64
4	70	58
4.5	63	52
5	39	60

The descriptive statistics in Table 2.5 below suggests that the independent task is easier than the integrated task but longer. There is more variance in scores in the integrated task, but the variance in text length is higher in the independent task. There also seems to be little score change depending on the topics.

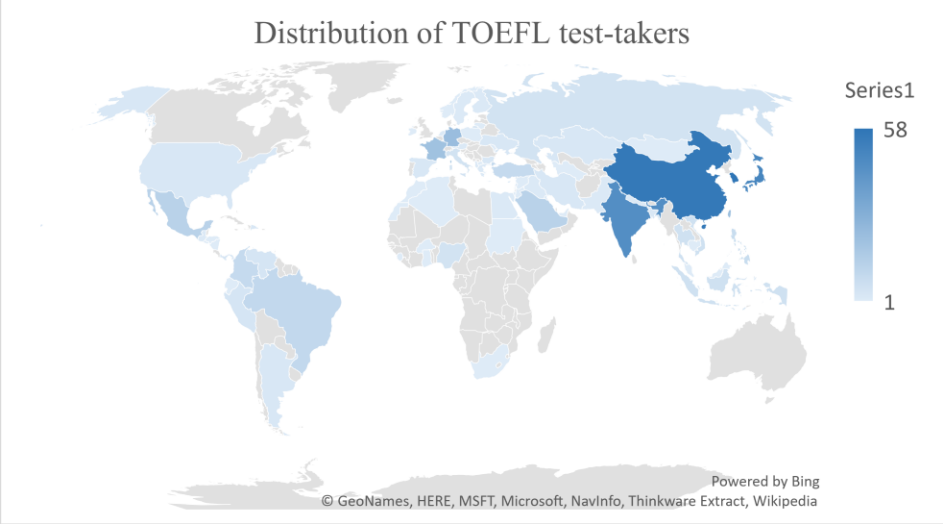
Table 2.5 Descriptive Statistics for the Integrated and Independent Writing Tasks

Task	*Topic*	*Scores (Mean)*	*Scores (SD)*	*Word Count (Mean)*	*Word Count (SD)*	*Word Count (Min)*	*Word Count (Max)*
Independent	Cooperation	3.47	0.91	310.60	76.70	86.00	586.00
	Career choice	3.38	0.86	324.70	79.30	61.00	558.00
Integrated	Bird migration	3.15	1.18	206.70	51.90	45.00	372.00
	Fish farming	3.15	1.31	196.80	50.60	54.00	388.00

2.5.4.1.3 Participants

The dataset was collected by ETS and the demographic information was provided for the 480 test-takers. The test-takers came from 76 different countries, the majority being from South Korea (*N* = 58), China, (*N* = 56), Japan, (*N* = 50), India (*N* = 46), Germany (*N* = 23), Taiwan (*N* = 22), and France (*N* = 21). Figure 2.2 shows the distribution of test-takers by country. There was a total of 52 first languages, which were predominantly Chinese (*N* = 83), Korean (*N* = 56), Spanish (*N* = 52), Japanese (*N* = 50), Arabic (*N* = 30), German (*N* = 26), French (*N* = 23), Hindu (*N* = 13), Russian (*N* = 10), and Portuguese (*N* = 10). The complete list of languages is provided in Appendix B. There were 204 females and 212 males (64 did not report gender), whose ages ranged from 14 to 51 years (mean = 23.6, SD = 6.4).

Figure 2.2 Distribution of Test-Takers by Country for the TOEFL iBT Public Use Dataset



2.5.4.1.4 Index Calculation for Essays

For the development of the L2 writing proficiency models, average frequency and range scores were computed for TOEFL essays from the independent (N = 480) and integrated task (N = 480). This step involved the computation of frequency and range scores for each lemma and n-gram lemma in the TOEFL essays that were available in the EF-CAMDAT and COCA Fiction. The frequency and range scores for the individual lemmas and n-grams were averaged for each essay, forming one frequency and one range score for each text and each index. TAALES (Kyle et al., 2018) was used to derive the COCA Fiction frequency and range scores for each text.

Most of the lemmas (97.8%) and bigrams (83.8%) in the TOEFL essays were assigned a score from EF-CAMDAT. Trigrams had a coverage of 48%, which is acceptable given the odds of matching three-lemma combinations. This coverage suggests that the EF-CAMDAT indices can provide output for several lemmas and n-gram lemmas present in L2 writing.

2.5.4.2 Lexical Decision Data

For the development of the L2 lexical processing models, reaction time and accuracy scores from a lexical decision task by Berger, Crossley, and Skalicky (2019) were used as outcome variables. The L2 behavioral data in Berger and colleagues come from an online crowdsourcing study that collected data from a lexical decision task (i.e., a word/non-word decision task) that included 3,318 English content words and 3,318 pseudowords judged by 1,315 self-identified L2 users of English. A summary of Berger and colleagues' data collection procedures is provided below.

2.5.4.2.1 Lexical Decision Task

A lexical decision (LD) task aims at testing lexical processing through a visual word recognition task which, along with word naming tasks, has been one of the "gold standards" in

developing models of word processing (Balota et al., 2007). In this task, participants are presented with a string of letters (i.e., the stimulus) and asked to press a button judging whether the stimulus is a word or a non-word. Accuracy data (i.e., correct or incorrect judgement of the stimulus) and reaction time (i.e., the time elapsed from the presentation of the stimulus and the response) are the standard measures of lexical processing. High accuracy rates and fast responses are indicative of higher entrenchment of lexical items in the mental lexicon (Brysbaert et al., 2017; Diependaele et al., 2013).

2.5.4.2.2 Participants

The online crowdsourcing platform Amazon Mechanical Turk, which allows web users to perform online tasks in exchange for financial compensation, was used to gather data from L2 users. Berger, Crossley, and Skalicky (2019) screened participants by applying a background questionnaire, which included questions about the languages the participants spoke and their dominant language. A total of 765 males and 550[7] females performed the task, and the majority reported using English more than four hours a day (*N* = 704, 54%), followed by those who used English 3–4 hours a day (*N* = 195, 15%), 2–3 hours a day (*N* = 168, 13%), 1–2 hours a day (*N* = 170, 13%), and less than one hour a day (*N* = 78, 6%). The majority of the participants were very confident (*N* = 595, 45%) to somewhat confident (*N* = 572, 43%) in their use of English. Most spent more than eight years studying English (*N* = 754, 57%), followed by those who studied English for 6–8 years (*N* = 214, 16%), 4–6 years (*N* = 168, 13%), 2–4 years (*N* = 140, 11%) and less than one year (*N* = 38, 3%). A total of 1,152 participants (88%) reported having lived primarily in an English-speaking country. Their length of residence varied substantially. A total of 239 participants (19%) reported having lived over 120 months in an English speaking country,

[7] Age is not reported in Berget et al. (2019) due to an issue with the task set-up.

383 participants (30%) reported living in an English speaking country for 37–120 months, 307 participants (24%) for 13–36 months, and 344 participants (27%) for 0–12 months. The participants reported a total of 79 dominant languages, with Spanish (*N* = 532, 40%) , English (*N* = 202, 15%), French (*N* = 66, 5%), German (*N* = 57, 4%), and Chinese (*N* = 49, 4%) being the most common.

2.5.4.2.3 Stimuli

Berger, Crossley, and Skalicky (2019) selected 3,318 pseudowords and 3,318 content words from the English Lexicon Project developed by Balota et al. (2007), a multi-university project that collected word-information data from university students in the United States, primarily English monolinguals, for 40,481 words and 40,481 non-words. The stimuli in Berger et al. were distributed into 63 sub-lists with 84 to 104 stimuli each. A minimum of 20 observations was collected per word and pseudoword, including reaction time (in ms) and accuracy (in percentage). These data were computed with the Qualtrics Reaction Time Engine and Testable.

2.5.4.2.4 Procedure

Once participants agreed with the consent and had their qualifications checked, they proceeded to perform the task. For each trial of the task, participants saw a fixation point for 250ms followed by the stimulus word or pseudoword. Participants were required to press the letter “Q” for pseudoword and the letter “P” for words. A total of 136,360 observations were collected.

2.5.4.2.5 Reaction Time and Accuracy Mean Scores

This dissertation uses the average reaction time and accuracy information from the 3,318 words from Berger, Crossley, and Skalicky (2019). Berger and colleagues calculated the means

based on a two-step outlier identification. They first eliminated any reaction time equal or below 200ms or equal or above 3000ms. Then, they computed standard deviation and removed any word information per participant that were three SDs below or above the mean, which resulted in 127,533 observations. Mean scores for accuracy per word were calculated after the two-step outlier identification. Average scores for each word are available as supplementary material at https://doi.org/10.1017/S0272263119000019. Table 2.6 shows the average of the mean scores used in the dissertation.

Table 2.6 Average of Observations from the Lexical Decision Task by Berger, Crossley, and Skalicky (2019)

Variables	***Average of Observations***
Average of L2 RT mean	734.158
Average of L2 Accuracy mean	0.940

2.5.5 Statistical Analysis

The first validation step entailed the creation of writing quality models. Linear mixed-effects models were computed using the TOEFL scores as the outcome variable and the frequency and range indices as the fixed effects. Language was entered as a random effect. Separate models for each TOEFL task (i.e., independent and integrated) were run considering the evidence that both tasks elicit different types of discourse (Enright & Tyson, 2008), including differences related to lexis (Biber & Gray, 2013; Cumming et al., 2005). A total of six models were run: one independent model and one integrated model for each corpus (i.e., EF-CAMDAT and COCA Fiction) and one independent combined and one integrated combined with all indices. The models were compared to judge the explanatory power of the L2 norms.

Linear mixed-effects (LME) modeling was chosen in this first validation step because the TOEFL dataset contains information about a random population of language speakers from different countries. Language is a contextual variable that brings dependency to the data,

meaning that residuals will be correlated; therefore, language needs to be accounted for as a random effect. Besides, LMEs work similarly to multiple regressions in that predictor variables can be used as fixed effects, including categorical variables (Baayen et al., 2008). In this dissertation, the frequency and range indices were tested as fixed effects along with essay topic, gender, and age. The integrated environment for R (RStudio Team, 2016) was used for the statistical analyses, along with the following packages: *lme4* (Bates et al., 2015), which was used to calculate the LME models, *lmerTest* (Kuznetsova et al., 2015) and *MuMIn* (Nakagawa & Schielzeth, 2013), which were used to obtain *p* values and marginal and conditional R^2 values for the fixed-effects model (i.e., the part of the model with fixed effects) and random-effects model (i.e., the part of the model with random effects) respectively. The *r2glmm* package (Jaeger, 2016) was used to calculate semi-partial R^2 for each fixed effect, which is an standardized measure of effect size. Note that due to differences in the marginal and semi-partial r-squared computations, the R^2 values for the fixed effects do not always sum up to be the same R^2 of the model.

The forward approach to model development was adopted. In this approach, an unconditional model with only by-L1 random intercepts was created (James et al., 2013; Murakami, 2016). Predictors (i.e., fixed effects) were added one by one, which were only kept if they decreased the AIC value (Akaike Information Criterion), which is used as a measure of model fit in comparison with similar models. Predictors with higher correlation coefficients with the outcome variable were added first. After the addition of each variable, the models were statistically compared using likelihood ratio tests, and the models with the best fit are reported. This approach allowed for full control of suppression effects and other issues with model development. Appendix C contains tables with the comparison statistics for the six LME models reported below.

Only the logarithmic transformed indices were tested in the models. This decision was based on the finding that low-frequency words (i.e., between 0.1 and 1 frequency per million words) from raw lists, which tend to compose 80% of corpora, show little predictive power (Heuven et al., 2014). The logarithmic transformation weighs the value of the lexical items in relation to the corpora, alleviating this issue. Additionally, logarithmic transformations make the distribution more linear, as opposed to Zipfian, which can result in more linear relationships with the outcome variables, a requirement for LMEs.

The first research question, which asked to which extent the L2 and L1 frequency and range indices predicted writing quality, was answered by checking the effect of the indices as fixed effects in the LME models. If the variables were significant predictors of writing quality and improved model fit, they were considered successful predictors. The models were also statistically compared using the r-squared difference test from the *r2glmm* package (Jaeger, 2016) to check whether there were statistically superior models. The combined model also provided information about the most predictive variables of writing quality; the significant predictors and the ones that explained more variance were considered more predictive.

The second validation step entailed the creation of lexical processing models. Frequency and range indices for all words from EF-CAMDAT and COCA Fiction were used as explanatory variables of reaction time and accuracy values from a lexical decision task performed by L2 users. For each word included in the lexical decision task, frequency and range scores were calculated. Linear multiple regression models were computed with reaction time and accuracy scores as the outcome variable using the *nlme* package (Pinheiro et al., 2017) in R. A forward and backward approach to model selection was adopted by using the function stepAIC() in R, which automatically performs model selection by comparing AIC values. The package *relaimpo*

(Grömping, 2006) was used to calculate the relative importance of each fixed effect in the multiple regression models. The LMG metric, which is the "R^2 contribution averaged over orderings among regressors" (Grömping, 2006, p. 13) is reported along with the marginal R^2 in the tables reporting model statistics. Two models per corpora were calculated: one with reaction time mean scores per word as the outcome variable and one with accuracy mean scores per word as the outcome variable. Combined models were also calculated. The lexical processing models were compared to judge the validity and predictive power of the L2 norms using the r-squared difference test.

The second research question, which asked to which extent the L2 and L1 frequency and range indices were predictive of lexical processing, was answered by checking the effect of the indices in multiple regression models. If the variables were significant predictors of reaction time and accuracy scores, they were considered successful predictors. The models were also statistically compared by using the r-squared difference test. If a model was statistically superior to the others, it was interpreted as an indication that its indices are stronger predictors of lexical processing.

2.6 Results

This results section is divided into two main parts: L2 writing quality models, which reports the models that explain the integrated and independent TOEFL scores, and lexical processing models, which reports the models explaining reaction time and accuracy from an LD task performed by L2 users.

2.6.1 L2 Writing Quality Models

The writing quality models section is divided into four parts: EF-CAMDAT models, COCA Fiction models, combined models, and model comparisons. For each corpus, correlations

between the frequency and range indices are provided, which is followed by the independent model and the integrated model.

2.6.1.1 EF-CAMDAT models

All EF-CAMDAT index scores for the TOEFL essays were checked for multicollinearity with a threshold set at $r \geq .7$. A few EF-CAMDAT frequency and range indices (see Table 2.3 for a complete list of the indices) were highly correlated with each other. The indices with higher correlations with writing scores and that were not highly correlated with other indices were kept. Table 2.7 below shows the correlation scores between the writing tasks and the selected indices. A dash ("–") indicates that the variable was multicollinear.

Table 2.7 Correlation Scores between the Dependent Variables and the Selected EF-CAMDAT Indices

EF-CAMDAT Frequency and Range Indices	***Independent Scores***	***Integrated Scores***
EF-CAMDAT Range – Lemma bigrams Log	–0.282***	–
EF-CAMDAT Frequency – Content Lemmas Log	–0.264***	–
EF-CAMDAT Frequency – Lemma trigrams Log	–	–0.227***
EF-CAMDAT Range – All Lemmas Log	–	–0.110*
EF-CAMDAT Frequency – Lemma bigrams Log	–	–0.147**
EF-CAMDAT Frequency – Function Lemmas Log	0.102*	–
EF-CAMDAT Range – Function Lemmas Log	–	–0.092*

*** p < .0005, ** p < .005, * p < 0.05

2.6.1.1.1 EF-CAMDAT Independent Model

The independent task model shows the effect of the frequency and range indices on the independent task scores. Task topic (i.e., a categorical variable with two values: cooperation, career choice), age, and gender (i.e., a categorical variable with two values: male and female) were used as control variables. Language (i.e., L1-background) was used as a random effect and the frequency and range indices as fixed effects. Table 2.8 shows the statistics for the

independent EF-CAMDAT model with the best fit, along with the semi-partial r-squared and 95% confidence intervals for each fixed effect.

Table 2.8 EF-CAMDAT Independent Model with Best Fit

Random Effects	***Variance***	***SD***					
Language (Intercept)	0.103	0.322					
Residual	0.609	0.780					
Fixed Effects	***Estimates***	***SE[a]***	***t-value***	***p***	***R^{2b}***	***95% CI***	
(Intercept)	3.276	0.291	11.239	<.005	0.11	0.16	0.06
EF-CAMDAT Contextual – Diversity of Lemma Bigrams Log	–0.738	0.096	–7.668	<0.05	0.10	0.16	0.06
EF-CAMDAT Frequency – Function Lemmas Log	1.145	0.355	3.225	<0.05	0.02	0.05	0.00

[a] Standard Error; [b] Marginal R^2 for the model and semi-partial R^2 for fixed effects

The fixed effects of the EF-CAMDAT model explained 11% of the variance (marginal R^2= 0.11), and L1-background explained 24% of the variance (conditional R^2= 0.24). Two variables were significant predictors of writing quality as measured by the independent task: range of lemma bigrams and frequency of function lemmas. Lemma bigrams explained most of the variance (10%), as shown by the semi-partial R^2. Age, gender, and topic were not significant control predictors of writing quality (see Appendix C for model comparison statistics) and were not entered in subsequent independent models.

2.6.1.1.2 EF-CAMDAT Integrated Model

The integrated task model shows the effect of the frequency and range indices on the integrated task scores. Task topic (i.e., a categorical variable with two values: bird migration and fish farming), age, and gender (i.e., a categorical variable with two values: male and female) were used as control variables. Language (i.e., L1-background) was used as a random effect, and the frequency and range indices as fixed effects. Table 2.9 shows statistics for the integrated EF-CAMDAT model with the best fit.

Table 2.9 EF-CAMDAT Integrated Model with Best Fit

Random effects	***Variance***	***SD***					
Language (intercept)	0.134	0.366					
Residual	1.352	1.163					
Fixed effects	***Estimates***	***SE***	***t-value***	***p***	***R^2***	***95% CI***	
(Intercept)	0.291	0.658	0.442	0.66	0.04	0.08	0.01
EF-CAMDAT Frequency – Lemma trigrams Log	–0.673	0.146	–4.620	<0.05	0.04	0.08	0.01

The only significant fixed effect in the integrated EF-CAMDAT model (i.e., frequency of lemma trigrams) explained 4% (marginal R^2= 0.04) of the scores, and L1-background explained 13% of the scores (conditional R^2= 0.127). Age, gender, and topic were not significant control predictors and were not added to subsequent integrated models.

2.6.1.2 COCA Fiction Models

The COCA Fiction index scores for the TOEFL essays were checked for multicollinearity with a threshold set at r ≥ .7. Table 2.10 shows the non-multicollinear indices and correlations with the dependent variables.

Table 2.10 Correlation Scores between the Dependent Variables and Selected COCA Fiction Indices

COCA Fiction Frequency and Range Indices	***Independent Scores***	***Integrated Scores***
COCA Fiction Range – Content Lemmas Log	–0.360***	–
COCA Fiction Frequency – Lemma Bigrams Log	–0.156***	–0.130*
COCA Fiction Frequency – Lemma Trigrams Log	–0.108*	–0.100*

*** p < .0005, ** p < .005, * p < 0.05

2.6.1.2.1 COCA Fiction Independent Model

An independent model similar to the EF-CAMDAT independent model was run with the COCA Fiction frequency and range indices. Table 2.11 shows the statistics for the independent model with the best fit.

Table 2.11 COCA Fiction Independent Model with Best Fit

Random effects	***Variance***	***SD***						
Language (intercept)	0.131	0.362						
Residual	0.585	0.765						
Fixed effects		***Estimates***	***SE***	***t-value***	***p***	***R^2***	***95% CI***	
(Intercept)		6.330	0.328	19.276	<.005	0.12	0.18	0.08
COCA Fiction Range – Content Lemmas Log		–17.164	1.973	–8.699	<.005	0.12	0.18	0.08

The fixed effect in the COCA Fiction independent model (i.e., range of content lemmas) explained 12% of the scores (marginal R^2= 0.12), and L1-background explained 24% of the scores (conditional R^2 = 0.24).

2.6.1.2.2 COCA Fiction Integrated Model

An integrated model similar to the EF-CAMDAT integrated model was run with the COCA Fiction frequency and range indices. Table 2.12 shows the statistics of the integrated model with the best fit.

Table 2.12 COCA Fiction Integrated Model with Best Fit

Random Effects	***Variance***	***SD***						
Language (intercept)	0.149	0.386						
Residual	1.389	1.179						
Fixed Effects		***Estimates***	***SE***	***t-value***	***p***	***R^2***	***95% CI***	
(Intercept)		4.863	0.613	7.931	<.005	0.01	0.04	0.00
COCA Fiction Frequency - Lemma Bigrams Log		–1.065	0.414	–2.572	0.01	0.01	0.04	0.00

The only significant fixed effect in the COCA Fiction Integrated model (i.e., frequency of lemma bigrams) explained 1% of the scores (marginal R^2 = 0.01), and L1-background explained 10% of the scores (conditional R^2 = 0.10).

2.6.1.3 Combined Models

All frequency and range scores from EF-CAMDAT and COCA Fiction for the TOEFL essays were checked for multicollinearity with a threshold set at $r \geq .7$. Table 2.13 shows the non-multicollinear indices and the correlations with the independent variables.

Table 2.13 Correlation Scores between the Dependent Variables and EF-CAMDAT and COCA Fiction Selected Indices

All Frequency and Range Indices	***Independent***	***Integrated***
COCA Fiction Range – Content lemmas log	–0.360***	–
EF-CAMDAT Range – Lemma bigrams log	–0.282***	–0.145**
EF-CAMDAT Frequency – Lemma trigrams Log	–	–0.227***
EF-CAMDAT Range – All lemmas log	–	–0.159*
COCA Fiction Frequency – Lemma trigrams log	–0.109*	–0.100*
EF-CAMDAT Frequency – Function lemmas log	0.102*	–

*** p < .0005, ** p < .005, * p < 0.05

2.6.1.3.1 Combined Independent Model

An independent model with all selected frequency and range indices from the two corpora was run. Table 2.14 shows the combined independent model with the best fit, along with the semi-partial r-squared and 95% confidence intervals for each fixed effect.

Table 2.14 Combined Independent Model with Best Fit

Random Effects	***Variance***	***SD***					
Language (intercept)	0.128	0.358					
Residual	0.581	0.762					
Fixed Effects	***Estimates***	***SE***	***t-value***	***p***	***R^2***	***95% CI***	
(Intercept)	4.034	1.073	3.76	<.005	0.13	0.19	0.09
COCA Fiction Range – Content Lemmas Log	–13.475	2.561	–5.262	<.005	0.05	0.09	0.02
EF-CAMDAT Range – Lemma bigrams Log	–0.276	0.123	–2.246	0.02	0.01	0.03	0.00

The fixed effects in the combined independent model explained 13% of the scores (marginal $R^2 = 0.13$), and the L1-background explained 27.5% of the scores (conditional $R^2 = 0.275$). One COCA Fiction (i.e., range of content lemmas) explained 5% of the scores and one

EF-CAMDAT index (i.e., range of lemma bigrams) explained 1% of the writing scores, as informed by the semi-partial r-squared.

2.6.1.3.2 Combined Integrated Model

An integrated model with all selected frequency and range indices was run. The model with the best fit was the same as the EF-CAMDAT integrated model reported in Table 2.9 above.

2.6.1.4 Model Comparisons and Research Questions

The models from the three corpora and combined models were statistically compared using the r-squared difference test. The independent model comparisons are provided in Table 2.15, and the integrated model comparisons are presented in Table 2.16. The indices included in each model and the percentage that each index explains is also provided.

Table 2.15 Comparisons between the EF-CAMDAT Independent Model and the COCA Independent Models

Independent Models	***Marginal R^2***	***AIC***	***Indices***	***Semi-partial R^2***	***EF-CAMDAT Independent***
EF-CAMDAT Independent	11%	1165.2	EF-CAMDAT Range – Lemma Bigrams Log	10%	–
			EF-CAMDAT Frequency – Function Lemmas Log	2%	
COCA Fiction Independent	12%	1151.4	COCA Fiction Range – Content Lemmas Log	12%	$r = –0.02$, $p = 0.32$
Combined Independent	13%	1148.4	COCA Fiction Range – Content Lemmas Log	5%	$r = –0.03$, $p = 0.24$
			EF-CAMDAT Range – Lemma bigrams Log	1%	

Table 2.16 Comparisons between the EF-CAMDAT Integrated Model and the COCA Integrated Models

Integrated Models	***Marginal R^2***	***AIC***	***Indices***	***Semi-partial R^2***	***EF-CAMDAT Integrated***
EF-CAMDAT Integrated	4%	1537.2	EF-CAMDAT Frequency – Lemma Trigrams Log	4%	–
COCA Fiction Integrated	1%	1551.5	COCA Fiction Frequency – Lemma bigrams Log	1%	$r = –0.027$, $p = 0.13$
Combined Integrated	–	–	Same as EF-CAMDAT Integrated	–	–

Research question number one asked to what extent the L2 and L1 indices explained L2 writing quality. As reported above, both EF-CAMDAT and COCA Fiction indices were successful predictors of writing quality, with no differences across models in terms of how much the models explained. However, there was an advantage for a COCA index in the combined independent model and, in the integrated combined model, only an EF-CAMDAT index surfaced as a predictor (i.e., frequency of trigrams). Overall, a combination of all different types of indices (e.g., content lemmas, bigrams) helped explain writing quality as measured by the independent task. The integrated models, on the other hand, preferred the n-gram indices. There was also an overall preference for range over frequency indices, but they were highly correlated, as multicollinearity analyses revealed. All indices but the function lemmas indices had a negative relationship with essay scores, meaning that when test takers used lemmas or n-gram lemmas that were less frequent, their scores were higher.

2.6.2 Lexical Processing Models

This section reports on the lexical processing models that explain reaction time and accuracy scores for the 3,318 words from Berger, Crossley, and Skalicky (2019). Models with

frequency and range of all words indices from EF-CAMDAT and COCA Fiction were used. All word indices instead of all lemma indices were preferred because morphemes play a role in lexical decision task performance (Muncer et al., 2014). Because the frequency and range indices were highly correlated, resulting in only one explanatory variable per outcome variable, only correlations are reported for the EF-CAMDAT models and COCA Fiction corpora[8], along with the R^2. However, a combined RT and a combined accuracy model are reported. The correlations between the selected indices and the dependent variables and r-squared values are provided in Table 2.17.

Table 2.17 Correlations between the RT and Accuracy Scores and the EF-CAMDAT and COCA Fiction Indices

Indices	***L2 RT***	***R^2***	***L2 Accuracy***	***R^2***
EF-CAMDAT Range – All Words Log (N = 3381)	–0.368***	0.135	0.374***	0.139
COCA Fiction Range – All Words Log (N = 3316)	–0.389***	0.151	0.337***	0.114

*** p < .0005, ** p < .005, * p < 0.05

The EF-CAMDAT range index explained 13.5% of the RT and 14% of the accuracy scores, and the COCA Fiction range index explained 15% of the RT and 12 % of the accuracy scores, as revealed by the r-squared values.

2.6.2.1 Combined Models

Multiple regression models were run combining the EF-CAMDAT and COCA Fiction all word indices as explanatory variables of RT and accuracy (Degrees of freedom = 3380), as reported in Tables 2.18 and 2.19.

[8] Linear regressions with only one variable provide the same results as correlations.

Table 2.18 Combined Reaction Time Model

Fixed Effects	*Estimates*	*SE*	*t-value*	*p*	R^{2a}
(Intercept)	635.151	4.275	148.585	< .005	0.17
COCA Fiction Range – All Words Log	–42.354	3.691	–11.476	< .005	0.09
EF-CAMDAT Range – All Words Log	–7.085	0.845	–8.384	< .005	0.08

[a] Adjusted R^2 for the model and LMG (i.e., R^2 partitioned) for predictors.

Table 2.19 Combined Accuracy Model

Fixed Effects	*Estimates*	*SE*	*t-value*	*p*	R^2
(Intercept)	1.032	0.004	257.826	< .005	0.15
EF-CAMDAT Range – All Words Log	0.009	0.001	11.974	< .005	0.09
COCA Fiction Range – All Words Log	0.022	0.003	6.451	< .005	0.06

The combined RT model explained 17% of the variance in RT scores. The COCA Fiction index explained more variance than the EF-CAMDAT index as suggested by the LMG value (i.e., R^2 partitioned). The combined accuracy model explained 15% of the variance, with the EF-CAMDAT index explaining more of the accuracy scores.

2.6.2.2 Model Comparisons and Research Questions

The correlation scores revealed that both EF-CAMDAT and COCA Fiction indices explained a similar amount of word processing scores. The EF-CAMDAT indices had an advantage in explaining accuracy scores, whereas the COCA Fiction indices had an advantage in explaining the RT scores. The combined models confirmed this trend.

Research question number two asked to what extent the L2 and L1 indices explained L2 lexical decision reaction time and accuracy scores. As reported above, the EF-CAMDAT and COCA Fiction indices were successful predictors of both reaction time and accuracy. The range indices had a higher correlation with the dependent variables and were all successful predictors of lexical processing. As expected, range had a negative relationship with reaction time, meaning that words that appear in more texts are named faster (i.e., have a lower RT value). Also as

expected, range had a positive relationship with accuracy, suggesting that words that appear in more texts are named more accurately. It is worth noting, though, that range and frequency indices were highly correlated. Overall, the statistical comparisons suggested that the models were compatible and that the indices are better seen as complementing each other.

2.7 Discussion

Lexical sophistication has been investigated in several L2 writing (e.g., Biber & Gray, 2013; Kyle & Crossley, 2016; Römer, 2009b) and L2 lexical processing studies (e.g., Brysbaert et al., 2017; Dijkstra & Heuven, 2002; Van Wijnendaele & Brysbaert, 2002). It has been a tradition in many of these studies to use corpus-based benchmarks derived from L1 corpora to assess L2 lexical proficiency. However, scholars have advocated for the use of L2 benchmarks as more direct representations of the L2 experience with language (Naismith et al., 2018; Vaid & Meuter, 2017). Finding and testing indices that more closely represent the linguistic experience of language users has been one of the major challenges in lexical proficiency research (Heuven et al., 2014), with scholars suggesting testing alternative benchmarks to reach a more accurate representation of exposure and sophistication (Adelman et al., 2006; Bestgen, 2017; Heuven et al., 2014). Building on this assumption, Study 1 of this dissertation tested the validity of frequency and range indices based on L2 corpora as representations of L2 lexical sophistication.

The first step in the validation of the L2 indices involved the use of these norms as explanatory variables of L2 writing, which replicates past research that has predominantly used L1 indices (e.g., Garner et al., 2019; Guo et al., 2013; Kyle & Crossley, 2016). TOEFL essays, which have also been extensively used in the L2 writing literature (e.g., Biber & Gray, 2013; Enright & Tyson, 2008; Friginal et al., 2014; Guo et al., 2013; Kyle et al., 2016), were selected for replication purposes. Both integrated and independent essays were included to test the indices

as explanatory variables of two distinct writing tasks. The first research question was addressed in the first validation step. This question asked to what extent the L2 and L1 lexical frequency and range indices were predictive of L2 writing proficiency. The models of L2 writing proficiency suggested that the L2 (i.e., EF-CAMDAT) and the L1 indices (i.e., COCA Fiction) explained a similar amount of essay score variance, with a slight advantage for the COCA indices when explaining the independent scores and a slight advantage for the EF-CAMDAT indices in explaining integrated scores; however, no statistical differences between the models were found.

A combination of lemma bigrams, content lemmas, and function lemmas were predictors of independent essay scores, whereas only n-gram indices explained integrated essay scores. The presence of more sophisticated bigrams (i.e., bigrams with lower frequency) from EF-CAMDAT was associated with higher scores in the independent task, and the presence of more sophisticated trigrams and bigrams from EF-CAMDAT and COCA Fiction was associated with higher integrated scores. This finding does not replicate the findings of previous statistical models of L2 writing, which have found that less sophisticated n-grams led to higher essay scores both when L1-based indices (e.g., Kyle et al., 2016; Kyle & Crossley, 2015) and L2-based indices (Monteiro et al., 2020) were used as benchmarks. However, it supports research that has found that proficient writers use more sophisticated lexical bundles (e.g., Ädel & Erman, 2012; Shin et al., 2018). The presence of more sophisticated content lemmas, as indexed by COCA Fiction, also led to higher scores, a relationship that replicates previous findings (Crossley et al., 2015; Guo et al., 2013; Kyle & Crossley, 2015; Palfreyman & Karaki, 2019). Finally, the presence of less sophisticated function words was associated with higher essay scores, but this index had a small impact in the assessment of writing quality, as previous research had already

suggested (Kyle et al., 2016; Kyle & Crossley, 2015). The indices COCA Fiction lemma bigrams and EF-CAMDAT function lemmas did not contribute to the combined models, suggesting that they are weak predictors in the presence of other indices.

An analysis of sample texts was performed to illuminate these findings. For illustrative purposes, Appendix D features the individual output of a high-scored and a low-scored independent essay, and Appendix E features a high-scored and low-scored integrated essay. Index scores for all significant indices are included, and a few non-significant indices were featured for comparisons between the L1 and L2 indices, which is performed in Chapter 5. The items that contributed to higher scores were highlighted in red; that is, values above or below the mean of all test takers were highlighted, depending on the index relationship with scores. For example, if the index had a negative relationship with essay scores, the items below the test takers' mean (i.e., the mean for the entire population included in this study) for that index were foregrounded. The appendices include the original text, a table with index scores for the selected essays, and tables containing the individual output for types (i.e., the unique lemmas in the text). Token count is also provided in the tables. The same procedure is adopted in Study 2 and 3.

Regarding the independent essays, an investigation of individual output as measured by the index EF-CAMDAT range bigram revealed that both high scorers and low scorers used a diversity of bigrams with phrasal verbs, noun phrases, prepositional phrases, and adverb phrases; however, high-scored essays contained more bigrams with adjectives and more sophisticated adverbs and nouns. In the example in Appendix D, some of the highly sophisticated bigrams used in the high-scored essay included "have fortunately," "very likely," "highly value(d)," "high demand," "initial goal," "financial independence," and "continue(ing) education." The low-scored essay contained bigrams with only two adjectives (i.e., "important" and "easy") and

more common adverbs such as “so” and “ago.” The effect of COCA Fiction content lemmas is similar to the effect of the bigrams; more sophisticated content lemmas led to higher scores. High scorers used substantially more sophisticated content lemmas (see example in Appendix D) including content lemmas that had very low range scores such as “frugal” and “self-realization.” The use of function words, as measured by the index EF-CAMDAT frequency of function lemmas, showed the opposite trend in terms of sophistication in that higher scores were associated with more common function words. The individual output in Appendix D clarifies this effect. While the high-scored essay contained more sophisticated function words such as “further” and “while” that were not present in the low-scored essay, the high-scored essay is much longer and elaborated, demanding the use of highly frequent articles such as “the” and “a” to specify noun phrases and the verb “to be” in copula and passive construction, all of which are highly frequent function words (see token count for “the,” “be,” and “a” in the high-scored essay). This trend was found in several other texts.

Integrated essays were also analyzed to understand the effect of trigrams and bigrams on essay scores. Both high-scored and low-scored essays contained topic-related word combinations that had low-frequency scores, as indexed by EF-CAMDAT; that is, they were considered sophisticated combinations. Such is the case of “pork and beef,” “contaminated by the,” and “by the chemical” taken from the high-scored essay featured in Appendix E, and “farm be (is) not,” “health due to,” and “have less fat” from the low-scored essay. However, the high-scored essays contained several other sophisticated trigrams, especially referential three-word combinations such as “the claim of,” “by the professor,” “the professor who,” “conjurer state that,” “the author argue(d),” which are important to structure an integrated essay. The same pattern was found for the bigrams indexed by COCA Fiction; both high scorers and low scorers used relatively

sophisticated topic-related bigrams such as “substance that,” “pound of,” “of commercial,” and “risk to” (see low-scored essay in Appendix E), but high scorers included more of bigrams that helped them report and organize the ideas from the sources such as “passage be(is),” “consider to,” “hence the,” and “hint that.”

Overall, the results from the L2 writing proficiency models strengthen the findings of previous research which suggests that lexical sophistication is an important component of L2 writing competency, with more proficient writers using more sophisticated words and n-grams (Crossley et al., 2015; Guo et al., 2013; Kyle & Crossley, 2015; Palfreyman & Karaki, 2019). The fact that frequency explained only a fraction of the writing scores is unsurprising considering the many discourse features that are related to writing quality (Biber & Gray, 2013; Cumming et al., 2006). Besides, due to the holistic nature of scoring, lexical sophistication can be less relevant if other writing quality criteria are met such as cohesion, completeness of the response, and appropriateness of argumentation (Biber & Gray, 2013; Jarvis et al., 2003). Finally, raters may be affected by how linguistic features are combined in a way that multiple successful profiles are possible (Friginal et al., 2014; Jarvis et al., 2003), making it difficult to tease out single linguistic features that can explain writing quality. The fact that the integrated essay scores had a lower variance explained by the indices (i.e., the frequency indices explained only 1% to 4% of the scores) replicates previous research (Guo et al., 2013; Kyle et al., 2016) and can be related to the nature of the task and the rubric. The presence of sources provides writers with sophisticated lexical items that interfere with the automatic investigation of the writer’s own lexical knowledge; that is, the integrated words have a confounding effect in the analysis. Also, while the rubric for the independent task makes explicit mention of “lexical errors” and “appropriate word choice,” the integrated task leaves those out in favor of content

and organization.

The second step in the validation of the L2 indices involved the use of these benchmarks as explanatory variables of L2 lexical processing, which replicates past research that has predominantly used L1 indices (e.g., Brysbaert et al., 2017; Diependaele et al., 2013; Duyck et al., 2008; Lemhöfer et al., 2008; Van Wijnendaele & Brysbaert, 2002). This step answered the second research question, which asked to what extent the L2 and L1 frequency and range indices for all words are predictive of L2 lexical decision reaction time and accuracy scores. Both word frequency and word range indices were considered, but because they were multicollinear and range indices had higher correlations with word processing measures, only range indices were tested in the models. The results suggested that the EF-CAMDAT index had a slight advantage in explaining accuracy, whereas the COCA Fiction index had a slight advantage in explaining RT scores (see Chapter 5 for a discussion of this trade-off effect). The differential effect of the indices was not tested statistically as only correlations were reported. This study has also found that the combined models explained more variance (i.e., 17% of RT scores and 15% of accuracy scores), suggesting that the indices were complementing each other.

The effect of range as indexed by both EF-CAMDAT and COCA Fiction was the same: higher frequency of texts led to more efficient word processing (i.e., faster and more accurate word judgments). Appendix F features the 100 words that were processed faster and more accurately by L2 users of English, and the 100 words that were processed more slowly and less accurately. Similar to the individual output from the independent and integrated essays, the values that facilitated processing are highlighted in red. There is a clear concentration of highlighted items (i.e., words with higher range) among the words that are processed faster and more accurately. Words that appear in more texts such as “couch,” “music,” “public,” and

"express" were processed more efficiently than words such as "sine," "tinker," "gram," and "grocer" that had a lower range score as indexed by both EF-CAMDAT and COCA Fiction. Chapter 5 discusses the exceptions and differences found for both corpora.

Overall, the results from the L2 word processing models strengthen the findings of previous research, which suggests that range impacts lexical processing, probably more so than word frequency (Berger, Crossley, & Skalicky; 2019; Hamrick & Pandža, 2020; Johns et al., 2012; Skalicky et al., in press). Words that appear in more texts were processed faster and judged more accurately, a finding that supports the hypothesis that repeated exposure across contexts strengthens the representation of lexical items in the L2 mental lexicon (Adelman et al., 2006). The results also support the "principle of likely need," which establishes that words that are needed in more contexts develop stronger representations in the mental lexicon (Jones et al., 2017). This effect is related to the constant activation of lexical items in multiple encounters with input, which work to strengthen the representations of these items in the mental lexicon (Jones et al., 2017).

2.8 Conclusion and Limitations

Study 1 suggested that frequency and range indices based on L2 corpora can be successfully used in the assessment of lexical proficiency. The results showed that while the L1 indices explained more of the independent essay scores, the L2 indices explained more of the integrated scores, and while the L1 indices explained more of the reaction time, the L2 indices explained more of the accuracy scores from the lexical decision task. This suggests that complementing text analyses with multiple corpora that represent the multiple linguistic varieties L2 users are exposed to have the potential of augmenting explanatory power, strengthening and broadening past conclusions regarding L2 production and processing.

The differential effect of the L2 and L1 indices open opportunities for future investigations regarding raters' expectations towards lexical sophistication. The results suggested that n-gram indices from EF-CAMDAT were more predictive of essay scores, whereas single-lemma indices from COCA Fiction were more predictive. Most of the n-grams that were indexed by EF-CAMDAT as more sophisticated were also indexed as more sophisticated by COCA Fiction; however, many n-grams that were indexed as less sophisticated by EF-CAMDAT were indexed as more sophisticated by COCA Fiction, including "last century," "a(n) external," "example of," "a(n) excellent," "job description," and "more important" (see Appendix D), which are seemingly common in academic and classroom writing. It is possible that raters judge n-gram sophistication based on the experience that L2 users may have when learning English, making the EF-CAMDAT n-gram indices more relevant for L2 text analysis. However, because the development of vocabulary knowledge may be easier than the development of phraseological knowledge (Ellis, 2002; Pawley & Syder, 1983), raters might consider native-like lemma knowledge when judging essays. This may explain the higher impact of single-lemma COCA Fiction indices. This hypothesis can be tested in future research through an investigation of raters' cognitions regarding sophistication judgements of lemmas and lemma n-grams that are indexed differently by L1 and L2 corpora.

Some limitations should be noted. Firstly, Study 1 was limited to two metrics of lexical sophistication (i.e., lexical frequency and range) and treated lexical sophistication as detached from grammar. Other measures, such as lexical density, diversity, and phrase frames, must be considered for a complete understanding of lexical proficiency. Also, the corpora used here may not fully represent the linguistic experience that the TOEFL test takers and the participants in the lexical decision task had. The TOEFL test takers, for example, were possibly exposed to

different varieties of English, topics, and genres that were not covered either by EF-CAMDAT and COCA Fiction; therefore, the corpora used in the present study are only a partial proxy of the language to which the L2 users were exposed. Regarding the lexical decision models, individual differences such as age and first language were not controlled for as they were in the L2 writing models. This was not possible with the use of average RT and accuracy scores per word as outcome variables. The lack of confounding fixed effects was another shortcoming of the models. Because the purpose here was to compare similar variables from different corpora, confounding variables such as semantic variables were not included to test the predictive power of the frequency and range indices in the presence of other lexical sophistication variables. Finally, the analysis of lexical sophistication in the integrated essays without controlling for the integrated words from the source may have resulted in inexact findings. As shown in the individual output, many of the lexical items in the essays were incorporated from the input texts. The individual scores from these items influenced the strength of the indices, primarily weakening them. An approach similar to the one adopted in Kyle (2020), where he analyzed the successful use of words from the sources, may have been more appropriate to the analysis of integrated essays.

This study has also brought to light some important considerations regarding the use of L2 corpora. As pointedly stated by Meurers and Dickinson (2017), automatic text analyses are not free of error and are dependent on important decisions related to text analyses, especially when it comes to L2 language, which is highly variable. One important consideration for frequency lists is cut-off points or the minimum frequency allowed in a list. Establishing a conservative cut-off point of raw frequency of five and above can eliminate half to two-thirds of words and word combinations, reducing the number of items automatically analyzed in a text.

However, a low cut-off point of two allows the inclusion of misspellings with low-frequency scores, which can lead to a few incorrect individual item scores. Such was the case of the bigram “with othe” in the low-scored independent essay in Appendix D, which had a low-frequency match in EF-CAMDAT. However, a cut-off point of five would have missed 25 bigrams in the high-scored essay in Appendix D. An informed decision such as the one adopted in this study is necessary. To test the most reliable cut-off point, models with a conservative approach (i.e., a cut-off point set at 5) and a less conservative approach (i.e., a cut-off point set at 2) were compared (not reported here due to space constraints), and the differences between the conservative and less conservative models were not significant. This may suggest that for a robust analysis with multiple texts, a comprehensive coverage with a low cut-off point may be acceptable. However, for other studies where a high level of precision is needed and comprehensive coverage is optional, a more conservative approach must be considered.

Another consideration when dealing with L2 corpora is the presence of highly frequent non-standard production. For example, the verb “belive” (“believe” in Standard English) is among the top 2,000 words in EF-CAMDAT. While spellcheckers can change non-standard forms to standard forms, the inclusion of non-standard instances may be relevant for studies of English as a foreign language or English as a lingua franca. Another issue with non-standard forms is faulty lemmatization. Forms like “belive" are not lemmatized because lemma lists are designed for standard language. These issues highlight the importance of individual output and qualitative analyses, as well as the development of new approaches to the automatic analysis of L2 production (Meurers & Dickinson, 2017).

3 STUDY 2: DEVELOPING AND TESTING L2 SEMANTIC CONTEXT INDICES

Recent developments in lexical processing research have established that the mental lexicon is a complex network of interconnected lexical items (Wilks & Meara, 2002; Zareva, 2007). Although many factors influence the architecture of lexical representations such as perceptual experience (e.g., sight and smell) and phonology, meaning is probably the strongest force structuring the mental lexicon (Landauer & Dumais, 1997; Lund & Burgess, 1996). Meaning-based theories of lexical acquisition argue that related lexical items are stored together and that most of these relationships are based on the analysis of the distribution of lexical items in the input (Landauer & Dumais, 1997; McDonald & Shillcock, 2001). Evidence supporting this meaning-based view comes from different sources. Corpus-based computational models, which simulates the architecture of the mental lexicon by modeling word co-occurrence, has corroborated the importance of semantics in structuring the mental lexicon (Jones et al., 2012; Landauer & Dumais, 1997; Lund & Burgess, 1996; Mikolov, Chen, et al., 2013). There is also psycholinguistic evidence from word recognition tasks with L1 (Balota et al., 2004; Jones et al., 2012; Lund & Burgess, 1996; McDonald & Shillcock, 2001) and L2 users (Berger, Crossley, & Skalicky, 2019; Johns et al., 2016; Skalicky et al., in press) that semantics influences word processing. Overall, these studies have suggested that semantic context benchmarks are stronger predictors of processing than lexical frequency and that the semantic distributional properties in the input may be what drives the frequency effect (McDonald & Shillcock, 2001).

Despite evidence that semantic context may be a more valid representation of the experience with language input, frequency-based benchmarks such as lexical frequency and range have been the norm in L2 writing studies (Crossley & McNamara, 2012; Johnson et al., 2016; Kyle & Crossley, 2015; Laufer & Nation, 1995; Meara & Bell, 2001). One of the reasons

for the widespread use of frequency norms is the abundant availability of automatic frequency indices and the limited availability of automatic semantic context indices. Another issue with the available indices developed so far is that they have been based primarily on L1 corpora, or corpora with edited texts such as Google News and TASA, which are limited in their representation of language use (McDonald & Shillcock, 2001). To address the lack of L2-based automatic semantic context indices and amplify the limited number of semantic context indices available, Study 2 of this dissertation set out to test semantic context indices developed from L2 corpora and two distributional computational methods: Latent Semantic Analysis (Landauer et al. 1998) and Word to Vector (Mikolov, Chen, et al., 2013).

Similar to Study 1, the validity of the L2 semantic context indices as measures of lexical proficiency that can be used in applied linguistics and psycholinguistic studies was tested in two steps. In step one, the indices were used to model L2 writing proficiency and, in step two, to model L2 lexical processing. These validations are meant to answer the second research question of this dissertation regarding the usefulness of the L2 semantic context indices as predictors of L2 writing and L2 word processing.

3.1 Semantic Context

Connectionist models of language acquisition are based on the premise that acquisition occurs from experience, with each event with language resulting in cognitive changes (Dell et al., 1999; Ellis, 2002). These models are patterned after computer models, representing an individual lexical item in a speaker's mind as a node that is connected to other related nodes as a function of linguistic experience (Dell et al., 1999). In the previous chapter, an argument was made that repeated experience, represented by lexical frequency and range, can strengthen lexical representations and entrench the connections among these lexical nodes (Adelman et al.,

2006; Balota & Chumbley, 1985; Ellis, 2002). One major criticism of this frequency effect is that language users do not experience lexical items discretely, but in a semantic environment in which lexical items are strongly related; therefore, a semantic account to explain lexical knowledge may be more appropriate (Lund & Burgess, 1996; McDonald & Shillcock, 2001).[9]

The idea that semantic context has a major impact on lexical proficiency is based on the premise that words that occur together share semantic similarities and that the experience with these semantically related items results in them being stored together (Landauer, 2007; McDonald & Shillcock, 2001). In the connectionist model analogy, the nodes representing semantic-related words are stored in proximity and have stronger connections. Several psycholinguistic studies have tested whether semantic co-occurrence is a force driving lexical proficiency, with most suggesting that semantic context indices are more reliable predictors of lexical processing than frequency, and are, therefore, a more valid representation of lexical entrenchment (Berger, Crossley, & Skalicky, 2019; Johns & Jones, 2008; Jones et al., 2012; McDonald & Shillcock, 2001; Skalicky et al., in press); however, this evidence is restricted to a limited number of words that can be subjected to human judgements. Computational models solve this problem by simulating lexical acquisition through the analysis of word co-occurrence from large corpora.

Distributional semantic models (DMS) use large corpora to simulate how humans use the statistical properties of language to represent word meaning (Jones et al., 2012; Landauer & Dumais, 1997; Lund & Burgess, 1996; Mikolov, Chen, et al., 2013). DMSs are based on the

[9] An effect of frequency cannot be discarded, though. Semantic context measures highly correlate with frequency measures, a phenomenon attributed to high semantic context words being needed more frequently (McDonald & Shillcock, 2001). Also, as argued by Ellis (2002), there is a multiplicity of elements interacting with frequency, making it difficult to tease out one single variable that explains complex phenomena such as lexical acquisition.

assumption that word meaning is dictated by the contexts of word usage (Cruse, 1986; Firth, 1957); therefore, a representative corpus can provide the statistical experience that humans have with language. Spatial DMSs represent word co-occurrence in vector spaces, which replicate how lexical items are distributed in the mental lexicon (Jamieson et al., 2018). Several computational methods have been used to develop DMSs, including Latent Semantic Analysis (Landauer, 2007), Latent Dirichlet Allocation (Blei et al., 2003), and Word to Vector (Mikolov, Chen, et al., 2013), to name only a few. The first evidence of the success of these models is that they are capable of obtaining vector spaces with semantically related words (Landauer, 2007; Mikolov, Chen, et al., 2013). The validity of these models has also been extensively tested against behavioral data (Mandera et al., 2017; Riordan & Jones, 2011), which have shown that DMSs can replicate human knowledge in many tasks, including in multiple-choice vocabulary tests (Landauer & Dumais, 1997), disambiguation of meaning tasks (Jamieson et al., 2018, Mandera et al., 2017), taxonomic classification tasks (Jamieson et al., 2018), and syntactic and semantic questions (Mikolov, Chen, et al., 2013). The success of DMSs in modeling human learning behavior has led to its use in assessing L2 writing and understanding L2 lexical processing behavior, as detailed below.

3.1.1 Semantic Context and L2 Writing

In the L2 writing literature, semantic information has been primarily used as a measure of semantic cohesion. In these studies, a DMS is developed from a large corpus and used to estimate the similarity of meaning between parts (e.g., between sentences, paragraphs, utterances) of an input text. Semantic similarity is a significant predictor of L2 writing quality, with proficient writers developing texts with parts that are more semantically related (Crossley, Kyle, et al., 2014; Guo et al., 2013); however, semantic similarity has also been found to be high

in lower-level writing (e.g., Bestgen et al., 2010; Foltz, 2007). Comparisons of L2 and L1 writing have also shown that L2 writers score higher in cohesion measures (Green, 2012). Bestgen et al. (2010) argue that the use of repeated words (i.e., lower lexical diversity) in L2 writing, especially lower-level L2 writing, might inflate cohesion measures quantified automatically with DMSs, explaining the high coherence found in less proficient writing.

Semantic information as measures of lexical sophistication has taken many forms. Lexical sophistication indices such as hypernymy, concreteness, imageability, semantic co-referentiality, meaningfulness, and polysemy, which represent semantic properties of words, have been successfully used in the investigation of L2 writing (Crossley et al., 2010; Crossley & McNamara, 2012). These studies have found that L2 writers move from less sophisticated to more sophisticated words that are less concrete (Crossley et al., 2015), less imageable (Crossley, Kyle, et al., 2014), more polysemous (Kyle et al., 2018), have fewer superordinate items or items that are more specific (Kyle et al., 2018), and have fewer associations with other words (Crossley & McNamara, 2012). Only recently, semantic context information has been incorporated into the analysis of lexical sophistication, but they have been applied mostly to the investigation of L2 speaking (Berger et al., 2017; Crossley et al., 2013). Berger et al. (2017), for example, used associative context indices (i.e., the number of associations a word has with other words) derived from behavioral data from association tasks, a contextual distinctiveness index from McDonald and Shillcock (2001), and a semantic ambiguity index from Hoffman et al. (2013). The word context indices explained 49% of the variance in human ratings of lexical proficiency in speaking, with more proficient speakers using more ambiguous and more distinct words. To the author's knowledge, the only study which included semantic context as a measure of lexical sophistication to explain L2 writing was Skalicky et al. (2019), who used an LSA-based semantic

context measure as an explanatory variable of human scores of creativity in L2 writing. This index had a significant and negative correlation with creativity scores, meaning that creativity was related to the use of words that are associated with less distinct contexts.

3.1.2 Semantic Context and L2 Word Processing

Lexical processing studies have used a plethora of semantic property measures such as concreteness and imageability as explanatory variables of word processing measures. Most of the evidence from semantic variables have come from L1 studies (e.g., Bates et al., 2001; Brysbaert et al., 2000; Cuetos & Barbón, 2006; Morrison et al., 2002). However, evidence suggesting that semantic variables affect L2 lexical processing also exists. Studies have found that more concrete words (Skalicky et al., in press), more imageable words, words that are present in more contexts, and words that are more accurately defined (de Groot et al., 2002)[10] are processed faster by bilinguals. No significant effect for semantic variables such as hypernymy in L2 lexical decision tasks has been reported (Berger, Crossley, & Skalicky, 2019; Hamrick & Pandža, 2020).

Measures that have used the distributional characteristics of words have also been tested in the lexical processing literature. Several methods have been used to develop these measures, including indices derived from psycholinguistic data such as word association norms from word association tasks (Kiss et al., 1973; Nelson & Friedrich, 1980), and corpus-based indices derived with computational methods (Hoffman et al., 2013; Johns et al., 2016; McDonald & Shillcock, 2001). Indices related to contextual distinctiveness (Berger, Crossley, & Skalicky, 2019; Skalicky et al., in press), word associations (Berger, Crossley, & Skalicky, 2019; Skalicky et al., in press), context availability (de Groot et al., 2002), and semantic diversity (Hamrick & Pandža,

[10] de Groot et al. (2002) used a semantic dimension, derived from PCA, as explanatory variable of lexical processing data. The semantic dimension included three indices: imageability, context availability, and definition accuracy.

2020; Johns et al., 2016) have surfaced as predictors of L2 lexical processing. Overall, these studies have shown that words that are related to more words and more contexts have a processing advantage. These studies also confirm that measures that account for the distributional nature of words are stronger predictors of processing than lexical frequency (Johns et al., 2016; Skalicky et al., in press), although range seems to explain processing beyond semantic context (Berger, Crossley, & Skalicky, 2019; Hamrick & Pandža, 2020; Skalicky et al., in press). These findings support the notion that contextual repetitions, modulated by semantic context, work to strengthen a word's memory (Jones et al., 2012), and highlight the importance of testing semantic context measures.

3.2 Research Questions

Study 2 was designed to answer the second research question of this dissertation regarding the predictive power of the L2 automatic semantic context indices as explanatory variables of L2 writing scores and L2 lexical processing data by themselves and in comparison with similar L1 automatic indices. The following specific research questions guided Study 2:

1) To what extent are L2 and L1 semantic context indices derived from written corpora predictive of L2 writing proficiency?
2) To what extent are L2 and L1 semantic context indices derived from written corpora predictive of L2 lexical decision reaction time and accuracy scores?

3.3 Methods

This study uses measures of semantic context as predictors of L2 writing quality and L2 word processing data from a lexical decision task. The semantic context indices were derived from the EF-CAMDAT corpus (Huang et al., 2017) using LSA (Landauer & Dumais, 1997) and Word2vec (Mikolov et al., 2013) computational methods. The L1 indices used for comparison

purposes were derived from the TASA corpus (Landauer, 2007) using LSA methods. The two distributional semantic models, EF-CAMDAT semantic context indices (i.e., L2 indices), TASA semantic context indices (i.e., L1 indices), dependent variables (i.e., TOEFL writing scores and lexical decision scores), and data analysis are outlined below.

3.3.1 Distributional Semantic Models

Both LSA and Word2vec are DMSs that generate a semantic space where words are represented by points (i.e., vectors), whose positions are dictated by the distributional properties of the words in a training corpus. The primary difference between both lies in the process used to generate the vector spaces. LSA calculates word relationships based on document boundaries by generating a term-document matrix which is decomposed using Singular Value Decomposition (SVD). In contrast, Word2vec works at the word level by collecting information from the surrounding words within a limited window size, which is fed to a neural network (R.-M. Botarleanu, personal communication, February 12, 2020). Details about model computations are provided below.

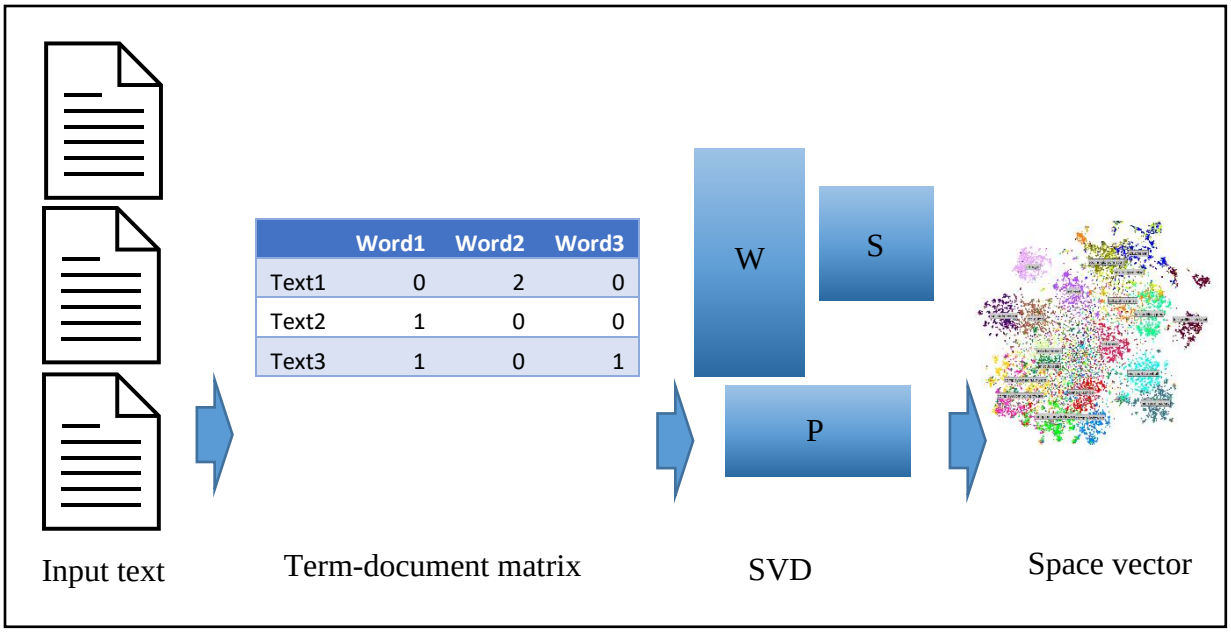
3.3.1.1 Latent Semantic Analysis

Latent Semantic Analysis is a technique for modeling word and text similarity. It takes as input a training corpus, which is transformed into a term-document matrix (i.e., a numeric representation of the distribution of words per text). A linear dimensionality reduction is applied, projecting the words in a multidimensional space. This process is detailed below.

The training corpus is usually preprocessed by eliminating function words and lemmatizing the words, increasing the distributional information per lemma (Riordan & Jones, 2011). The corpus information is organized in a term-document matrix with lemmas as columns and documents (e.g., texts) as rows. The resultant sparse matrix (i.e., a large matrix with mostly

zero values) is decomposed to generate word relationships. The decomposition of the term-document matrix is done using SVD, a method similar to Principal Component Analysis, which transforms the matrix into space vectors with the dimensions with the highest eigenvalues (Jamieson et al., 2018). The SVD decomposes the sparse matrix into three matrices that are truncated to reduce the number of rows and columns with little variance and multiplied back together, resulting in a more informative and reduced matrix (Lane et al., 2019). Co-occurrence among lemmas is computed by correlating the lemmas in the reduced matrix, while simultaneously finding the correlation between documents (i.e., texts) and documents and words. With the correlation results, linear combinations are created with related terms. These terms are represented in a vector space (i.e., a combination of vectors, each representing a different lemma) as dimensions. Figure 3.1 summarizes the LSA method.

Figure 3.1 Representation of LSA Method



3.3.1.2 Word2vec

Word2vec also represents words numerically into matrices that are factorized, but instead of working from a document-term matrix, vectors (i.e., a matrix with one column) are formed

locally. Therefore, while in LSA words that occur in the same document are treated as similar, in Word2vec, words must occur in proximity. Two neural networks are used to develop Word2vec: Continuous Bag of Words (CBOW), which predicts the word from context, and Skip-gram, which predicts the context words from a target word. The CBOW approach, used in this dissertation, is detailed below.

In the CBOW approach, a sliding window moves over every *n* words (e.g., every five words) in a corpus, creating vectors that feed the neural network. For example, for the sentence "NLP has helped many clients to make their life better," co-occurring information is gathered for the first five lemmas, for the second to the sixth, for the third to the seventh and so on, as represented in Figure 3.2. These iterations generate input vectors, also known as hot vectors. The hot vectors indicate the presence or absence of terms in a given sentence the same way that the term-document represents the presence of words in texts (Arumugam & Shanmugamani, 2018). The Softmax function, which calculates probabilities for a given set of values, is used to generate probabilities for the words in the neural network, which are used to map words into multidimensional vectors. Figure 3.3 provides a graphic representation of a neural network using the CBOW approach.

Figure 3.2 Representation of a Five-Word Rolling Window Centered at the Word "Clients"

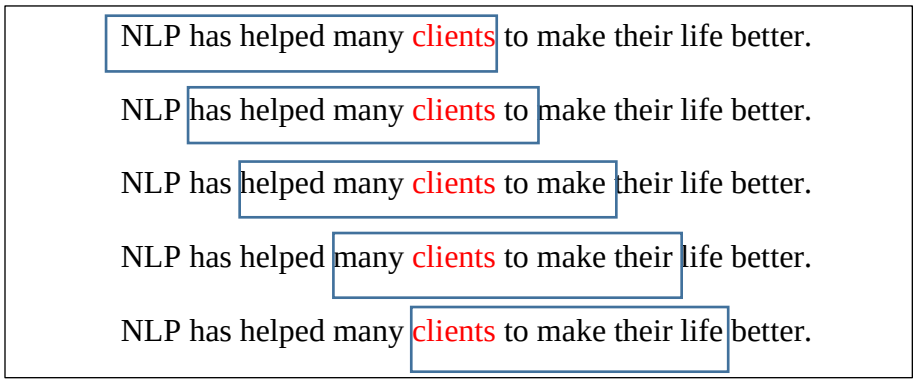
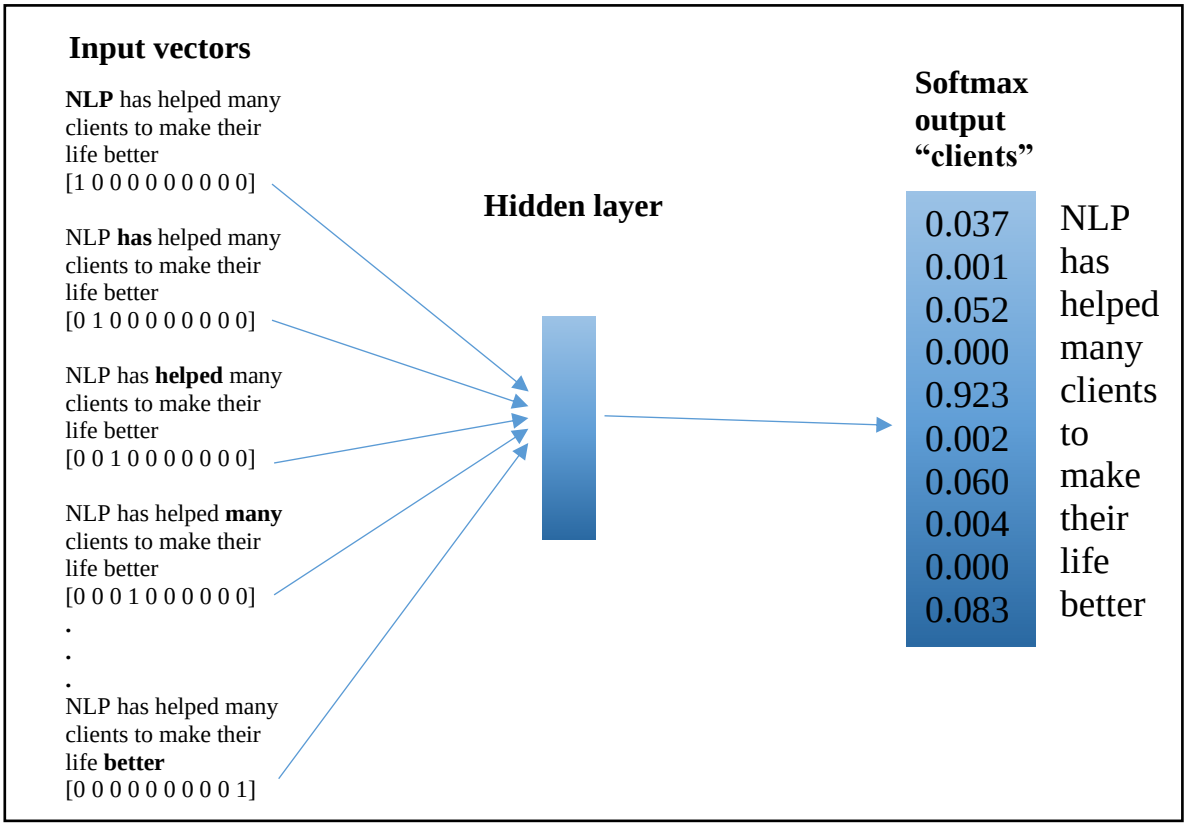


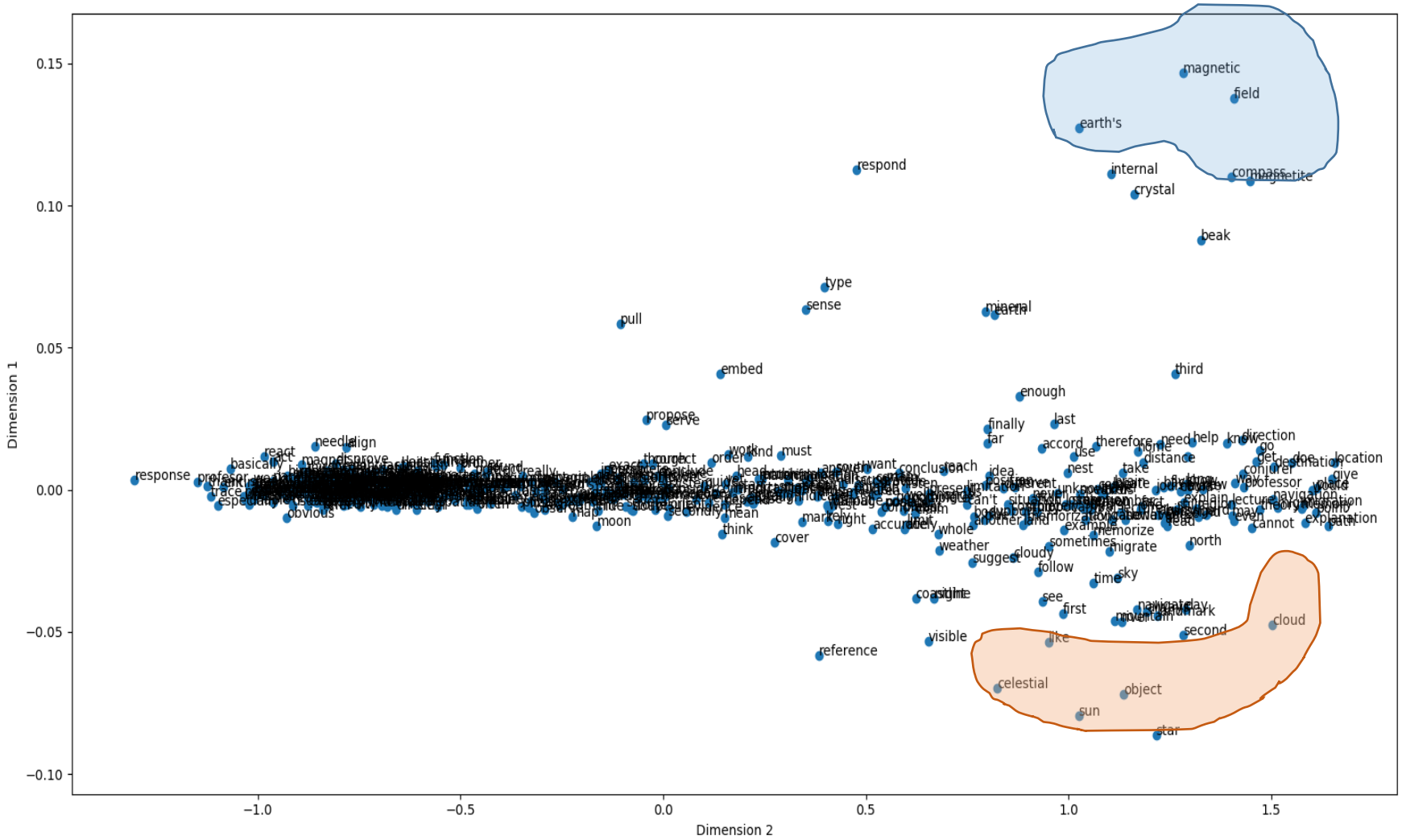
Figure 3.3 Representation of a CBOW Word2vec Network



3.3.1.3 Vector Space and Metrics

Vector spaces, also known as semantic spaces, are representations of the relationships between words and concepts generated by DMSs. The distribution of lemmas in the vector space represents the distribution of lemmas in the mental lexicon (Landauer, 2007). Hundreds of dimensions and millions of words are required to train models to obtain useful semantic information. Still, for illustrative purposes, a two-dimensional vector space using the Word2vec method was created with the integrated TOEFL essays on the topic of bird migration (N = 240). Figure 3.4 below shows the distribution of the words in a random vector space with two dimensions. Due to the low amount of texts, topics are not easily identifiable, but a few clusters emerged. The lemmas in the blue area are related to location/orientation, whereas the lemmas in the orange area are related to celestial objects.

Figure 3.4 Example of a Vector Space with Two Dimensions



To translate vector space information into benchmarks that can inform text analysis, models are trained incrementally (i.e., texts are added in batches), and model maturation is tested. This method is particularly helpful in finding information about lemmas that develop mature representations in intermediate or advanced models. A regression line is fit between intermediate and mature models, and the slope (i.e., the change in the y-coordinate divided by the change in the x coordinate) of the best fitting linear regression is used to judge lemma maturation (R. Botarleanu, personal communication, Feb 12, 2020). Lemmas with a lower slope develop representations early. Other useful information about lemmas is the distance between the vectors (i.e., lemmas). The metric used to calculate this distance is cosine similarity or the cosine of the angle between two vectors. A value close to 1 means the two vectors are close (i.e., the lemma pair is highly related). The number of cosines related to a lemma is another useful metric related to how semantically rich a lemma is. Thresholds for cosine values are established to only allow for meaningful relationships to be considered (Dascălu et al., 2016). For example, if lemma pairs below a cosine value of .3 are uninformative (i.e., the lemmas seem unrelated), a threshold of .3 and above is established. Examples for these measures (i.e., cosine-based indices, number of cosines, and slope) are provided in section 3.3.2.2 below.

3.3.2 EF-CAMDAT Indices

3.3.2.1 EF-CAMDAT Corpus

Similar to the frequency and range indices reported in chapter two, the semantic context indices were derived from the EF-CAMDAT corpus (Huang et al., 2017), which offers a range of topics and, because of size, is a good candidate to analyze word relationships (Lund & Burgess, 1996). The *Englishtown* levels seven to sixteen, which correspond to the levels B and C of the Common European Framework of Reference for languages (CEFR), were used. In total, the

levels contained 30,771,991 words from 246,328 texts. For more information about the corpus, refer to chapter two.

For both the LSA and Word2vec models, only lemmatized content words with a frequency of five and above were extracted. Potential misspellings in the corpus were removed by using the spellchecker library *Enchant.* This allowed the comparison of the content lemmas to a British and American dictionary as well as removed misspellings. Only the content lemmas that were judged to be an English lemma were entered in the analysis. Also, the vocabulary size for the LSA and word2vec models were limited to 2,000,000 lemmas, and 300 dimensions were allowed. All these measures were adopted to minimize the computational costs of calculating the models and reduce noise in the data. For Word2vec, the continuous bag of words (CBOW) approach was adopted with a window size set at five, and the *gensin* library (Rehurek & Sojka, 2010) was utilized for training both the LSA and Word2vec models in Python. Because cosines below .3 generated uninformative relationships, a threshold was set at .3. Semantic context information was calculated for 16,031 lemmas.

3.3.2.2 EF-CAMDAT Semantic Context Indices Selection

The following metrics of semantic context were used in this dissertation: cosine similarity scores, slope, and number of lemmas in the vector space. These indices are discussed in greater detail below.

3.3.2.2.1 Cosine Similarity Indices

Two types of cosine similarity indices were included in Study 2: highest cosines and average of cosines. Highest cosine similarity values were computed between all content lemmas in the corpus and the closest lemma (i.e., EF-CAMDAT – Highest cosine similarity index), the second closest lemma (i.e., EF-CAMDAT – Second highest cosine similarity index), and the

third closest lemma (i.e., EF-CAMDAT – Third highest cosine similarity index). Lemmas with a high value for cosine similarity have at least a few close associations with other words. For example, the word "porch" is highly associated with the word "staircase" (LSA highest cosine similarity = 0.9997) [11], "trim" (0.9994), and "remodel" (0.9992). The word "people," which appears in multiple contexts, has low cosine similarity values. The closest words to people are "unreceptive," "athleticism," and "amputee," with LSA cosine values of 0.27, 0.20, and 0.19, respectively. Correlations with other semantic or contextual benchmarks revealed moderate and positive relationships with concreteness and familiarity (see Appendix G for correlations between the LSA and Word2vec indices and other semantic variables), suggesting that lemmas with close associations with other lemmas may be more concrete and imageable. Moderate and positive correlations with EF-CAMDAT frequency and range indices also indicate that many lemmas with a high cosine value are relatively common. The highest cosine indices also highly correlated with each other (*r* ranged from .90 to .99). In sum, cosine indices indicate that a lemma occurs in a few distinct environments and may be used to index semantic distinctness.

Three types of average of cosine similarity values were calculated. These indices included the average top three cosines (i.e., EF-CAMDAT – Average top three cosines), the average of cosines above .3 threshold (i.e., EF-CAMDAT – Average cosine above .3 threshold), and the average of all cosines from intermediate and mature models (i.e., EF-CAMDAT – Average of all cosines). The index EF-CAMDAT – Average of the top three cosines is closely related to the top cosine indices described in the previous paragraph (i.e., correlations were positive and above .96). The index EF-CAMDAT – Average of cosines above .3 was also highly and positively correlated with top cosine indices (i.e., correlations ranged from .75 to .89). The

[11] For simplification purposes, the LSA lists were randomly selected as the baseline for exemplification.

threshold of .3 only allows distinct relationships to be included such that the higher the average of cosines following the .3 threshold, the more semantically distinct the lemmas are. Adjectives and adverbs such as "irresistible" (LSA average of cosines above .3 = .318) "shortly" (.307), and "definitely" (.308)" are among the lemmas with the lowest average above .3 cosine scores (i.e., they occur in less distinct environments), whereas specialized nouns such as "forerunner" (.711) and "fragmentation" (.716) are among the lemmas with highest average cosines above .3 because they occur in unique or distinct contexts.

The average of cosines above the .3 threshold had a non-significant correlation with average of all cosines (r = –0.03 for LSA, and r = – 0.04 for Word2vec), suggesting that these two indices are measuring different semantic relationships. As shown in Appendix G, the index EF-CAMDAT – average of all cosines was moderately and positively correlated with familiarity and meaningfulness, moderately and negatively correlated with age of acquisition, and highly and positively correlated with EF-CAMDAT frequency and range. This suggests that a lemma with a high average of cosine values tends to appear in several semantic contexts, including more distinct contexts, which increases their scores. For example, the lemmas "drama" (LSA average of all cosines = .718), "dessert" (.755), and "chicken" (.747), appear both in more restricted semantic contexts, but also in less restricted ones, reporting one of the highest average of all cosines scores. The lemmas with the lowest average cosine scores included highly specific lemmas such as "intelligentsia" (0.005), "triumphant" (0.005), and "bogeyman" (0.005). This index may be used to represent semantic richness for capturing the relationships among all related words.

In sum, for containing information about lemma relationships in restricted contexts, top-cosine indices (e.g., highest cosine similarity, average top 3 cosines) and the average of cosines

above .3 may serve to index semantic distinctiveness. The index average of cosines, on the other hand, takes into consideration all semantic relationships, being, therefore, representative of semantic richness.

3.3.2.2.2 Number of Cosines

The number of cosines in a vector space, with a threshold set at .3, was also added to the repertoire of LSA and Word2vec indices (i.e., EF-CAMDAT – Number of cosines above .3). The threshold of .3 only allows close and more distinct relationships to be embodied in this index. In other words, this index is representative of distinct relationships such that lemmas with a high number of cosines tend to be specific and with closely related lemmas. For example, the word "porch" reports a number of cosines score of 387, meaning that it is related to 387 words with a cosine of .3 and above. The word "people," on the other hand, has no related lemma with a cosine of .3 and above. This index behaves similarly to the top cosine indices in ranking semantically distinct lemmas (correlations between number of cosines and top cosines indices ranged from .45 to .62), but it seems to better capture the lemmas that occur in few distinct environments (i.e., the ones with fewer related lemmas) and in more distinct environments (i.e., the ones with more related lemmas). For this reason, this index may be used to capture distinctly rich lemmas.

3.3.2.2.3 Slope

Slope provides a measure of whether the lemma had mature representations in earlier models that included lower levels of EF-CAMDAT or in models where higher EF-CAMDAT levels were added. Lemmas that appear in earlier EF-CAMDAT models (i.e., in models with lower *Englishtown* levels) such as "school" (LSA slope = 0.000069), "Internet" (0.000072), and "dream" (0.0003) tend to have low slope values (i.e., they mature earlier). In contrast, lemmas

that converged later (i.e., they appeared in models with higher *Englishtown* levels) such as "felony" (0.148), "charisma" (0.162), and "expert" (slope = 0.141) have a higher slope. This measure may be used as a proxy of age of acquisition, or of lexical items that are more likely needed and, therefore, appear earlier in the corpus.

3.3.3 TASA Indices

The L1 semantic context indices available in TAALES were developed from TASA (Touchstone Applied Sciences Associates; http://lsa.colorado.edu/spaces.html) using LSA (Landauer et al., 1998). TASA has been widely used in cognitive science and educational research to represent average American college freshman students' reading experiences throughout their life (e.g., Dascălu et al., 2014, 2016; Johns & Jones, 2008; Landauer & Dumais, 1997). The TASA corpus is composed of texts ranging from 3rd grade to college level, and it includes a variety of genres such as novels, newspaper articles, samples from textbooks, and works of literature and fiction. Because the TASA indices represent classroom experience as does the EF-CAMDAT, they were considered strong candidates for comparisons. It is worth noting, though, that TASA represents classroom reading experience, whereas EF-CAMDAT represents classroom writing experience. Also, the TASA indices were developed in a similar fashion as the indices developed for this dissertation, and, to the authors' knowledge, no other compatible semantic context indices were freely available.

The indices available in TAALES were developed from 38,000 documents and 92,000 lemmas from TASA (Kyle et al., 2018). It should be noted that of the 38,000 documents, many of them are samples taken from the same text; therefore, the data are not independent. TAALES reports three TASA LSA indices, with a score provided for each lemma in the corpus. These indices are LSA contextual distinctness (maximum cosine), which is the cosine score for the top

related lemmas and the target lemma; LSA contextual distinctness (top 3 cosine), which is the average LSA cosine scores for the top three related lemmas; and LSA contextual distinctness (all cosine), which is the average of the LSA cosine scores for all related lemmas. Slope and number of cosine indices are not available in TAALES for L1 corpora. The computation of the TASA LSA indices approximates the computation of the indices developed for this dissertation, but computational details are not available for these indices. TAALES reports semantic context information for 4,487 lemmas.

3.3.4 Summary of Indices

The LSA and Word2vec indices from EF-CAMDAT and LSA TASA indices included in Study 2 are summarized in Table 3.1 below. The correlations between the TASA indices and the EF-CAMDAT are provided in Appendix G. The correlations coefficients were all small, suggesting the corpora were substantially different.

Table 3.1 List of LSA and Word2vec Indices from EF-CAMDAT and TASA

EF-CAMDAT LSA Indices	***EF-CAMDAT Word2vec Indices***	***TASA LSA Indices***
EF-CAMDAT LSA – Highest cosine similarity	EF-CAMDAT W2V -Highest cosine similarity	TASA LSA – Maximum similarity cosine
EF-CAMDAT LSA – Second highest cosine similarity	EF-CAMDAT W2V -Second highest cosine similarity	
EF-CAMDAT LSA – Third Highest cosine similarity	EF-CAMDAT W2V -Third Highest cosine similarity	
EF-CAMDAT LSA – Average top three cosine	EF-CAMDAT W2V -Average top three cosine	TASA LSA – Average of top 3 cosines
EF-CAMDAT LSA – Average cosine above .3	EF-CAMDAT W2V -Average cosine above .3	
EF-CAMDAT LSA – Average of all cosines	EF-CAMDAT W2V -Average of all cosines	TASA LSA – Average of all cosines
EF-CAMDAT LSA – Slope	EF-CAMDAT W2V -Slope	
EF-CAMTAT LSA – Number of cosines above .3	EF-CAMTAT W2V – Number of cosines above .3	

An interpretation of the indices as measures of lexical sophistication is provided in Table 3.2 below. The definitions and examples provided below represent tendencies and not absolute interpretations. Analyses of the indices and correlations with other semantic indices, which were, overall, low to moderate, suggest that they measure a range of semantic distributional behaviors. For example, lemmas with a highest cosine similarity score tend to be semantically distinct (see examples below); however, lemmas that behave similarly to other lemmas such as the verb "like," which is closely related to "want," have moderate to high cosine values as well, even though they may not be semantically distinct. Indices that only analyze top cosines or relationships above a certain threshold are also limited indicators of lemma relationships. For example, lemmas with a large high cosine value of .99 such as "departure," which is closely related to "airport," and "sore," which is closely related to "throat," also occur in less distinct environments as suggested by their high average of all cosines scores (i.e., .83 and .82, respectively). Finally, lemmas that are intuitively distinct such as "bazaar" and "piracy" have low cosine scores, which are indicative of low distinctness (i.e., it does not occur in distinct or unique contexts). These seemingly unexpected findings seem to be common among low-frequency lemmas, as it is the case of "bazaar" and "piracy," which only appear eight and six times on EF-CAMDAT respectively.

Table 3.2 Semantic Context Indices with Definitions and Examples

	Related Construct	***High scores***	***Examples***	***Low scores***	***Examples***
Highest cosines[a]	Semantic distinctness/ uniqueness	Lemmas that occur in distinct contexts	cardio, sodium, drain	Lemmas that occur in fewer distinct contexts	closely, head, vacate
Average of all cosines	Semantic richness	Lemmas with a rich network of semantic relationships	school, study, travel	Lemmas with a weak network of semantic relationships	aristocracy, bigfoot, bliss

Average of cosines above .3	Semantic distinctness/ uniqueness	Lemmas that occur in distinct contexts	decapitate, possibly, vacate	Lemmas that occur in fewer distinct contexts	internally, reopen, predominant
Number of cosines above .3 threshold	Semantic distinctness/ richness	Lemmas that occur in several distinct contexts	character, festival, slavery	Lemmas that occur in fewer distinct contexts	missionary, rapport, enumerate
Slope	Semantic maturation/ need	Lemmas that matured later; lemmas that are needed/acquired later	claustro-phobic, felony, butler	Lemmas that matured earlier; lemmas that are needed/ acquired earlier	robot, fuzzy, lice

[a] Highest cosine, second highest cosine, third highest cosine, and average of top 3 cosines

3.3.5 Outcome Variables

3.3.5.1 TOEFL Essay Scores

As in Study 1, the semantic context indices were used to develop models of L2 writing proficiency. Independent (N = 480) and integrated essays (N = 480) from the TOEFL iBT were analyzed using the L2 semantic context indices (i.e., EF-CAMDAT LSA and Word2vec indices) and similar L1 semantic context indices available in TAALES. The index scores were tested as predictors of the integrated and independent essay scores as judged by human raters. For more information about the TOEFL essays, refer to Chapter 2. The EF-CAMDAT indices covered 76% of the lemmas in the independent and integrated essays.

3.3.5.2 Lexical Decision Data

As in Study 1, models of lexical processing were developed using reaction time and accuracy data from a lexical decision task performed by L2 users from Berger, Crossley, and Skalicky (2019). The L2 and L1 semantic context indices were tested as predictors of L2 lexical processing. For more information about the lexical decision task, refer to Chapter 2.

3.3.6 Statistical Analysis

The data analysis for this study was similar to study 1. For the development of the L2 writing proficiency models, L1 (i.e., TASA) and L2 (i.e., EF-CAMDAT) semantic context indices were computed for the lemmas in the independent (N = 480) and integrated essays (N = 480). TAALES (Kyle et al., 2018) was used to compute the L1 semantic context indices. Linear mixed-effects models were calculated using the integrated and independent TOEFL scores as the outcome variables, language as a random effect, and the semantic context index average scores as fixed effects. The model comparisons from Chapter 2 suggested that none of the control variables (i.e., age, essay topic, gender) were significant fixed effects predicting either the integrated or the independent essay scores, as compared to the unconditional model. Therefore, for simplification purposes, the models that are built in this chapter exclude these control variables and only add language as a random effect, which had a strong effect.

Marginal and conditional r-squared for the models are provided, along with semi-partial r-squared for each fixed effect. Note that due to differences in the marginal and semi-partial r-squared computations, the r-squared values for the fixed effects do not always sum up to be the same as the model. The forward approach was adopted, and model comparisons statistics are provided in Appendix H. One independent and one integrated model for each corpus (i.e., EF-CAMDAT and TASA) were developed, as well as a combined independent and a combined integrated model. The models were statistically compared using the r-squared difference test.

The first research question, which asked to what degree the L2 and L1 semantic context indices were predictive of writing quality, was answered by checking the effect of the indices as fixed effects in the LME models. R-squared values and statistical comparisons were used as a measure of index and model effectiveness. Models and indices that explained more of the

variance in the writing scores were considered stronger predictors of L2 writing quality.

For the development of L2 lexical processing models, the same L1 and L2 indices were used as explanatory variables of reaction time and accuracy scores from a lexical decision task performed by L2 users. For each word included in the lexical decision task, L2 and L1 semantic context indices were calculated. Linear multiple regression models were computed for each corpus (i.e., TASA and EF-CAMDAT) and each outcome variable (i.e., reaction time and accuracy). Both a forward and backward approach to model selection were adopted by using the stepAIC() function. Degrees of freedom for the models and r-squared values (i.e., adjusted r-squared for the model and LMG for predictors) are included. To provide a comparison across available indices, all EF-CAMDAT (N = 16,031) and TASA lemmas (N = 4,487) were used to analyze the TOEFL essays and lexical decision words. Note that the combined RT and accuracy models only allow for overlapping items.

The second research question, which asked to what degree the L2 and L1 semantic context indices were predictive of lexical processing, was answered by checking the effect of the indices in multiple regression models. R-squared values were used to inform the strength of the indices. Indices with larger r-squared values were considered stronger predictors of L2 lexical processing.

3.4 Results

This results section is divided into two main parts: writing quality models and lexical processing models. The writing quality models section is further divided into four parts: EF-CAMDAT models, TASA models, combined models, and model comparisons. The lexical processing model section is divided into three parts: reaction time models, accuracy models, and model comparisons.

3.4.1 Writing Quality Models

3.4.1.1 EF-CAMDAT models

The EF-CAMDAT index scores for the TOEFL essays were checked for multicollinearity with a threshold set at $r \geq .7$. The indices with higher correlations with writing scores and that were not highly correlated with other indices were kept. Table 3.3 below shows the correlation scores between the writing tasks and the non-multicollinear indices. A dash ("–") indicates that the index was multicollinear.

Table 3.3 Correlation Scores between the Dependent Variables and the Selected EF-CAMDAT Indices

EF-CAMDAT Semantic Context Indices	***Independent Scores***	***Integrated Scores***
EF-CAMDAT LSA – Average of all cosines	–	–0.153***
EF-CAMDAT W2V – Average of all cosines	–0.427***	–
EF-CAMDAT LSA – Slope	–	0.171***
EF-CAMDAT LSA – Number of cosines above .3	0.352***	–
EF-CAMDAT LSA – Average cosine above .3	0.336***	–
EF-CAMDAT W2V – Number of cosines above .3	0.268***	0.160***
EF-CAMDAT W2V – Average cosine above .3	–0.232***	–
EF-CAMDAT W2V – Highest cosine similarity	–	–0.129**
EF-CAMDAT W2V – Slope	–	0.294***

*** $p < .0005$, ** $p < .005$, * $p < 0.05$, $p > .05$

3.4.1.1.1 EF-CAMDAT Independent Model

The EF-CAMDAT independent essay model shows the effect of the semantic context indices on the independent essay scores. Language was used as a random effect and the EF-CAMDAT semantic context indices as fixed effects. Table 3.4 below shows the independent EF-CAMDAT model with the best fit along with the r-squared values and 95% confidence intervals for each fixed effect.

Table 3.4 EF-CAMDAT Independent Model with Best Fit

Random effects	***Variance***	***SD***						
Language (intercept)	0.092	0.303						
Residual	0.558	0.747						
Fixed effects		***Estimates***	***SE[a]***	***t-value***	***p***	***R²[2b]***	***95% CI***	
(Intercept)		18.823	2.436	7.727	<.005	0.18	0.26	0.14
EF-CAMDAT W2V – Average of all cosines		–18.123	2.607	–6.950	<.005	0.09	0.14	0.04
EF-CAMDAT LSA – Number of cosines above .3		0.005	0.002	2.687	0.01	0.01	0.04	0.00

[a] Standard Error; [b] Marginal R^2 for the model and semi-partial R^2 for fixed effects

The fixed effects explained 18% of the scores (marginal $R^2 = 0.180$) and the random effects explained 29% of the scores (conditional $R^2 = 0.29$). The most significant predictor was the average of cosine similarities, which explained 9% (semi-partial $R^2 = 0.09$) of the independent scores, followed by number of cosines (1%).

3.4.1.1.2 EF-CAMDAT Integrated Model

The EF-CAMDAT integrated task model shows the effect of the EF-CAMDAT semantic context indices on the integrated task scores. Language was used as a random effect, and the EF-CAMDAT semantic context indices as fixed effects. Table 3.5 below shows the integrated EF-CAMDAT model with the best fit and its statistics.

Table 3.5 EF-CAMDAT Integrated Model with Best Fit

Random effects	***Variance***	***SD***						
Language (intercept)	0.146	0.382						
Residual	1.260	1.123						
Fixed effects	***Estimates***	***SE***	***t-value***	***p***	***R²***	***95% CI***		
(Intercept)	3.331	1.744	1.910	0.06	0.10	0.16	0.06	
EF-CAMDAT W2V – Slope	107.959	20.317	5.314	<.005	0.05	0.10	0.02	
EF-CAMDAT W2V – Number of cosines above .3	0.002	0.001	2.638	0.01	0.01	0.04	0.00	
EF-CAMDAT W2V – Highest cosine similarity	–6.347	2.710	–2.342	0.02	0.01	0.04	0.00	

The fixed effects explained 10% of the scores (marginal $R^2 = 0.100$) and the random effects explained 19.4% of the scores (conditional $R^2 = 0.194$). Slope was the predictor that explained most of this variance (5%, semi-partial $R^2 = 0.05$), followed by number of cosines above .3 (1%), and highest cosine (1%).

3.4.1.2 TASA Models

The three TASA index scores for the TOEFL essays were checked for multicollinearity with a threshold set at r ≥ .7. None of the TASA indices had a significant correlation with the independent essay scores, and only one TASA index had a significant correlation with the integrated scores, as shown in Table 3.6 below.

Table 3.6 Correlations Scores between the Dependent Variables and Selected TASA Indices

TASA LSA – Indices	***Independent***	***Integrated***
TASA LSA – Average all cosine	0.067	0.023
TASA LSA – Max similarity cosine	0.061	–
TASA LSA – Average top three cosine	–	–0.139**

*** $p < .0005$, ** $p < .005$, * $p < 0.05$, $p > .05$

3.4.1.2.1 TASA Independent Model

The TASA independent task model shows the effect of the TASA semantic context indices on the independent essay scores. Language was used as a random effect, and the TASA semantic context indices were entered as fixed effects. Despite the lack of significant correlations, the TASA indices were tested for comparison purposes, and, as expected, they did not make a significant contribution to the model and were eliminated, as shown in Table 3.7.

Table 3.7 TASA Independent Model with Best Fit

Random effects	***Variance***	***SD***		
Language (intercept)	0.126	0.355		
Residual	0.683	0.826		
Fixed effects	***Estimates***	***SE***	***t-value***	***p***
(Intercept)	3.545	0.077	46.34	<.005

The random effect of L1 background explained 16% of the scores (conditional $R^2 = 0.16$).

3.4.1.2.2 TASA Integrated Model

The TASA integrated essay model shows the effect of the TASA semantic context indices on the integrated task scores. Language was used as a random effect, and the TASA semantic context indices as fixed effects. Table 3.8 below shows the integrated TASA model with the best fit and its statistics.

Table 3.8 TASA Integrated Model with Best Fit

Random effects	***Variance***	***SD***					
Language (intercept)	0.156	0.395					
Residual	1.339	1.157					
Fixed effects	***Estimates***	***SE***	***t-value***	***p***	***R^2***	***95% CI***	
(Intercept)	4.166	0.648	6.432	<.005	0.05	0.09	0.02
TASA LSA – Average top three cosine	–20.802	4.321	–4.814	<.005	0.04	0.09	0.02
TASA LSA – Average all cosine	15.775	3.988	3.956	<.005	0.03	0.07	0.01

The fixed-effects model explained 5% of the scores (marginal $R^2 = 0.05$) and L1 background explained 14.5% of the scores (conditional $R^2 = 0.145$). Average of the top three cosines explained 4% (semi-partial $R^2 = .004$) of the scores, followed by the average of all cosines (3%).

3.4.1.3 Combined Models

All EF-CAMDAT and TASA index scores were checked for multicollinearity with a threshold set at $r \geq .7$. Table 3.9 below shows the correlation scores between the writing tasks and the non-multicollinear indices.

Table 3.9 Correlations Scores between the Essay Scores and All Semantic Context Indices

Indices	*Independent*	*Integrated*
EF-CAMDAT W2V – Average all cosines	–0.427***	–
EF-CAMDAT LSA – Number of cosines above .3	0.351***	–
EF-CAMDAT LSA – Average cosine above .3	0.336***	–
EF-CAMDAT W2V – Slope	–	0.294***
EF-CAMDAT W2V – Average cosine above .3	–0.232***	–
EF-CAMDAT W2V – Number of cosines above .3	0.268***	0.160***
EF-CAMDAT LSA – Slope	–	0.170**
EF-CAMDAT LSA – Average all cosines	–	–0.153***
TASA LSA – Average top three cosine	–	–0.139**
EF-CAMDAT W2V – Highest cosine word similarity	–	–0.129***
TASA LSA – Average all cosine	0.067	–
TASA LSA – Max similarity cosine	0.061	–

*** $p < .0005$, ** $p < .005$, * $p < 0.05$, $p > .05$

3.4.1.3.1 Combined Independent

After testing all EF-CAMDAT and TASA indices reported above, the combined model resulted in the same model as the EF-CAMDAT independent model reported in Table 3.4 above.

3.4.1.3.2 Combined Integrated

The combined integrated essay model shows the effect of the EF-CAMDAT and TASA semantic context indices on the integrated task scores. Table 3.10 below shows the integrated combined model with the best fit and its statistics.

Table 3.10 Combined Integrated Model with Best Fit

Random effects	*Variance*	*SD*					
Language (intercept)	0.143	0.379					
Residual	1.255	1.121					
Fixed effects	***Estimates***	***SE***	***t-value***	***p***	***R^2***	***95% CI***	
(Intercept)	–1.174	0.829	1.416	<.005	0.10	0.16	0.06
EF-CAMDAT W2V – Slope	112.160	19.825	5.658	<.005	0.06	0.11	0.03
EF-CAMDAT W2V – Number of cosines above .3	0.002	0.001	2.450	0.01	0.01	0.04	0.00
TASA LSA – Average top 3 cosine	–8.844	3.211	–2.754	0.01	0.01	0.04	0.00

The fixed-effects model explained 10.4% of the scores (marginal $R^2 = 0.104$) and the random effect (i.e., L1-background) explained 19.6% of the scores (conditional $R^2 = 0.196$). The index EF-CAMDAT Slope explained 6% (semi-partial $R^2 = .06$) of the scores, followed by number of cosines above .3 (1%) from EF-CAMDAT, and average of cosines above .3 threshold (1%) from TASA.

3.4.1.4 Model Comparisons and Research Questions

The EF-CAMDAT, TASA, and combined integrated models were statistically compared using the r-squared difference test. Because the TASA indices made no contributions to the independent scores, statistical comparisons were not performed between independent models. Table 3.11 summarizes the independent model statistics, and Table 3.12 shows the comparisons with the EF-CAMDAT integrated model. The fixed effects and percentage of variance explained by each index are also provided.

Table 3.11 Statistics for Independent Models

Independent Models	***Marginal and Conditional R^2***	***AIC***	***Indices***	***Semi-partial R^2***
EF-CAMDAT Independent	18%, 29%	1134.1	EF-CAMDAT W2V – Average of all cosines	9.00%
			EF-CAMDAT LSA – Number of cosines above.3	1.00%
TASA Independent	NA, 16%	1219.7	No significant fixed effects	NA
Combined Independent			Same as EF-CAMDAT Independent	

Table 3.12 Comparisons with the EF-CAMDAT Integrated Model

Integrated Models	***Marginal and conditional R^2***	***AIC***	***Indices***	***Semi-Partial R^2***	***EF-CAMDAT Integrated***
EF-CAMDAT Integrated	10%, 19.4%	1508.0	EF-CAMDAT W2V – Slope	5.00%	
			EF-CAMDAT W2V – Number of cosines above .3	1.00%	
			EF-CAMDAT W2V – Highest cosine similarity	1.00%	
TASA Integrated	5%, 14.5%	1536.4	TASA LSA – Average top three cosine	4.00%	$r = 0.054$, $p < .05$
			TASA LSA – Average all cosine	3.00%	
Combined Integrated	10.4%, 19.6%	1505.8	EF-CAMDAT W2V – Slope	6.00%	
			EF-CAMDAT W2V – Number of cosines above .3	1.00%	$r = –004$, $p = 0.45$
			TASA LSA – Average top three cosine	1.00%	

Tables 3.11 and 3.12 summarize the answer to research question one, which asked to what extent the L2 indices (i.e., EF-CAMDAT indices) and L1 indices (i.e., TASA indices) of semantic context explained writing quality. Only the EF-CAMDAT semantic context indices were significant predictors of the independent essay scores. None of the TASA LSA measures were significant predictors in the TASA independent model, and they did not contribute to the combined model (i.e., they did not improve the fit of the model and were, therefore, excluded). The integrated scores were explained both by the TASA and EF-CAMDAT semantic context indices, but the TASA model was statistically weaker than the EF-CAMDAT model. Also, only one TASA index was a significant predictor in the combined model. Regarding the effect of specific EF-CAMDAT indices, there was an overall preference for Word2vec indices, with average of all cosines being the best predictor in the independent model and slope being the best predictor in the integrated model. Slope was also a significant predictor of the independent scores, and number of cosines above .3 and highest cosine similarity contributed to the integrated

scores. Two TASA indices helped explain the integrated scores: average top three cosines and average of all cosines.

In sum, the answer to the first research question is that the L2 indices were stronger predictors of L2 writing, especially Word2vec indices. In the independent model, writers that gave preference to less semantically rich lemmas (i.e., lemmas with a weaker network of semantic relationships), but lemmas with a rich network of close relationships (i.e., lemmas that had rich and semantically distinct relationships), scored higher. In the integrated model, writers that gave preference to lemmas that develop representations later (i.e., they are learned later) and have a rich network of close relationships scored higher.

It is important to note that differences between TASA and EF-CAMDAT may not be related to L2 and L1 differences. TASA is composed of edited texts, whereas EF-CAMDAT is composed of student writing, which more closely resembles the TOEFL essays. Also, many more EF-CAMDAT indices were tested, increasing the chances of finding a better model (Murakami, 2016). Despite these limitations, for the same index types (e.g., average of all cosines), the EF-CAMDAT indices showed a much higher predictive strength.

3.4.2 Lexical Processing Models

Similar to Study 1, to test the power of the EF-CAMDAT semantic context indices, regression models were developed with reaction time and accuracy scores as dependent variables from a lexical decision task by Berger, Crossley, and Skalicky (2019). Different from Study 1, lemmas were investigated instead of words because the semantic context indices are only represented as lemmas. The words from the task were converted to lemmas, and the EF-CAMDAT and TASA indices were calculated. There was not EF-CAMDAT semantic context information for 170 out of the 3,318 words from Berger, Crossley, and Skalicky (2019), and 638

out of the 3,318 words were not available in TASA. The reaction time models, the accuracy models, and the combined models are reported below.

3.4.2.1 Reaction Time Models

The semantic context scores for the lexical decision words were checked for multicollinearity with a threshold set at $r \geq .7$. Table 3.13 shows the correlation scores for both the selected EF-CAMDAT and selected TASA indices with the RT scores.

Table 3.13 Correlation Scores between the RT Scores and Selected Semantic Context Indices.

Indices	***Accuracy Mean***
EF-CAMDAT W2V – Average of all cosines	–0.319***
EF-CAMDAT W2V – Number of cosines above .3	0.252***
EF-CAMDAT LSA – Highest cosine similarity	–0.090***
EF-CAMDAT W2V – Third Highest cosine similarity	0.063***
EF-CAMDAT LSA – Slope	0.057***
TASA LSA – Average all cosines	0.045*
EF-CAMDAT LSA – Number of cosines above .3	–0.024

*** $p < .0005$, ** $p < .005$, * $p < 0.05$, $p > .05$

3.4.2.1.1 EF-CAMDAT RT Model

A regression model was run with the EF-CAMDAT semantic context indices (degrees of freedom = 3,153) as explanatory variables of reaction time and is reported in Table 3.14.

Table 3.14 EF-CAMDAT RT Model with Best Fit

Indices	***Estimates***	***SE***	***t value***	***p***	***R^{2a}***
(Intercept)	761.760	3.436	221.704	<.005	0.086
EF-CAMDAT W2V – Average of all cosines	–0.025	0.001	–17.394	<.005	0.079
EF-CAMDAT W2V – Third highest cosine	0.009	0.001	5.965	<.005	0.006

[a] Adjusted R^2 for the model and LMG (i.e., R^2 partitioned) for predictors.

The model explained 8.6% of the variance (adjusted R^2 = 0.086). The index that explained most of the variance was the average of cosine similarities (8%, LMG = 0.079), followed by the third highest cosine similarity (1%).

3.4.2.1.2 TASA RT Model

Because only one TASA index was not multicollinear with the other indices, resulting in one fixed effect, only correlations are reported[12], along with the R^2. As shown in Table 3.14 above, TASA LSA – average of all cosine (N = 2,680) had a positive correlation with RT scores (r = 0.045), explaining less than 1% of the variance (R^2 = 0.002).

3.4.2.1.3 Combined RT Model

A regression model was run with the EF-CAMDAT and TASA semantic context indices as explanatory variables of reaction time (degrees of freedom = 2,580). The model with the best fit is reported in Table 3.15 below.

Table 3.15 Combined RT Model with Best Fit

Indices	***Estimates***	***SE***	***t value***	***p***	***R^2***
(Intercept)	751.329	4.363	172.191	<.005	0.096
EF-CAMDAT W2V – Average of all cosines	–0.024	0.002	–15.756	<.005	0.072
EF-CAMDAT LSA – Highest cosine similarity	–0.003	0.002	–2.044	0.040	0.003
EF-CAMDAT W2V – Third highest cosine similarity	0.008	0.002	4.934	<.005	0.005
TASA LSA – Average of all cosines	0.009	0.002	5.677	<.005	0.014

The combined RT model explained 9.6% of the variance (adjusted R^2 = 0.096). The index that explained most of the variance was the average of cosine similarities from EF-CAMDAT (7%, LMG = 0.072), followed by the average of all cosines from TASA (1%). The remaining indices explained less than 1% of the variance.

[12] Linear regressions with only one variable provide the same results as correlations.

3.4.2.2 Accuracy Models

The semantic context scores for the lexical decision words were checked for multicollinearity with a threshold set at $r \geq .7$. Table 3.16 shows the correlation scores for both the selected EF-CAMDAT and selected TASA indices with the accuracy scores.

Table 3.16 Correlations between Semantic Context Indices and Accuracy Scores

Indices	***Accuracy Mean***
EF-CAMDAT W2V – Average of all cosines	0.368***
EF-CAMDAT W2V – Number of cosines above .3	–0.252***
EF-CAMDAT LSA – Highest cosine similarity	0.148***
EF-CAMDAT LSA – Number of cosines above .3	0.053**
EF-CAMDAT W2V – Third Highest cosine similarity	–0.049**
TASA LSA – Average top three cosine	–0.062**
EF-CAMDAT LSA – Slope	0.019

*** $p < .0005$, ** $p < .005$, * $p < 0.05$, $p > .05$

3.4.2.2.1 EF-CAMDAT Accuracy Model

A regression model was run with the EF-CAMDAT semantic context indices (degrees of freedom = 3,143) as explanatory variables of accuracy and is reported in Table 3.17.

Table 3.17 EF-CAMDAT Accuracy Model with Best Fit

Indices	***Estimates***	***SE***	***t value***	***p***	***R^2***
(Intercept)	0.9094	0.004	215.355	<.005	0.085
EF-CAMDAT W2V – Average of all cosines	0.00002	0.011869	15.289	<.005	0.065
EF-CAMDAT W2V – Third highest cosine	–0.00001	0.000001	–6.899	<.005	0.008
EF-CAMDAT W2V – Number cosines above .3	–0.00001	0.000003	–2.288	0.020	0.001
EF-CAMDAT LSA – Highest cosine similarity	0.00001	0.000001	2.964	0.003	0.007
EF-CAMDAT LSA – Slope	0.00001	0.000001	3.0535	0.002	0.002

The model explained 8.5% of the variance (adjusted $R^2 = 0.085$). Average of all cosines was the best predictor, explaining 6.5% of the variance as suggested by the LMG value (i.e., R^2 partitioned). All other predictors explained less than 1% of the accuracy scores.

3.4.2.2.2 TASA Accuracy Model

Because only one TASA index was not multicollinear with the other indices, resulting in one fixed effect, only correlations are reported. The index TASA LSA – average of top three cosine scores (N = 2,680) had a negative correlation with accuracy (r = –0.062), explaining less than 1% of the scores (R^2 = 0.004).

3.4.2.2.3 Combined Accuracy Model

A regression model was run with the EF-CAMDAT and TASA semantic context indices (degrees of freedom = 2,573) as explanatory variables of accuracy and is reported in Table 3.18.

Table 3.18 Combined Accuracy Model with Best Fit

	Estimates	***SE***	***t value***	***p***	***R^2***
(Intercept)	0.920533	0.004650	197.958	<.005	0.093
EF-CAMDAT W2V – Average of all cosines	0.000022	0.000001	14.617	<.005	0.062
EF-CAMDAT W2V – Third highest cosine	–0.000009	0.000001	–6.134	<.005	0.007
EF-CAMDAT W2V – Number of cosines above .3	–0.000008	0.000003	–2.231	0.020	0.001
EF-CAMDAT LSA – Highest cosine	0.000005	0.000002	2.901	0.003	0.007
EF-CAMDAT LSA – Slope	0.000005	0.000002	3.047	0.002	0.001
TASA LSA – Average top three cosine	–0.000321	0.000058	–5.554	<.005	0.013

The combined model explained 9.3% of the variance (adjusted R^2 = 0.093). Similar to the EF-CAMDAT accuracy model, average of all cosines was the best predictor, explaining 6.2% of the accuracy scores as suggested by the LMG value. The remaining indices explained less than 1% of the variance in accuracy scores.

3.4.2.3 Model Comparisons and Research Questions

Table 3.19 shows the comparisons with the EF-CAMDAT RT model, and Table 3.20 shows the comparisons with the EF-CAMDAT Accuracy model. The fixed effects and percentage of variance explained by each model and index are also provided. Note that statistical

comparisons with the TASA models are not included because TASA models were not developed (i.e., only correlations were computed).

Table 3.19 Comparisons between RT Models

RT Models	***Adjusted R^2***	***AIC***	***Significant Indices***	***R^2***	***Comparisons***
EF-CAMDAT	8.6%	37433	EF-CAMDAT W2V – Average all cosines	8.00%	
			EF-CAMDAT W2V – Third highest cosine	0.06%	
TASA RT	NA	NA	TASA LSA – Average all cosine	0.20%	NA
Combined	9.6%	38400	EF-CAMDAT W2V – Average all cosines	7.20%	$r = –1.01$, $p = 0.23$
			EF-CAMDAT LSA – Highest cosine	0.30%	
			EF-CAMDAT W2V – Third highest cosine similarity	0.50%	
			TASA LSA – Average of all cosines	1.40%	

Table 3.20 Comparisons between Accuracy Models

Accuracy Models	***Adjusted R^2***	***AIC***	***Significant Indices***	***R^2***	***Comparisons***
EF-CAMDAT Accuracy	8.50%	–7806	EF-CAMDAT W2V – Average all cosines	6.50%	
			EF-CAMDAT W2V – Third highest cosine	0.80%	
			EF-CAMDAT W2V – Number cosines above .3	0.07%	
			EF-CAMDAT LSA – Highest cosine	0.70%	
			EF-CAMDAT LSA – Slope	0.20%	
TASA Accuracy	NA	NA	TASA LSA – Average Top Three Cosine	0.20%	NA
Combined Accuracy	9.30%	–7835	EF-CAMDAT W2V – Average of all cosines	6.20%	$r = –0.008$ $p = 0.26$
			EF-CAMDAT W2V – Third highest cosine	0.70%	
			EF-CAMDAT W2V – Number of cosines above .3	0.01%	
			EF-CAMDAT LSA – Highest cosine	0.70%	
			EF-CAMDAT LSA – slope	0.10%	
			TASA LSA – Average top three cosine	1.30%	

Research question two asked to what extent the L2 indices (i.e., EF-CAMDAT indices) and L1 indices (i.e., TASA indices) of semantic context explained lexical processing. Similar to the writing proficiency models, the L2 indices (i.e., EF-CAMDAT) were stronger predictors of word processing. None of the TASA indices contributed to the Combined RT model. The accuracy scores were also predominantly explained by the EF-CAMDAT semantic context indices, with a marginal contribution of the TASA index average top three cosines in the combined accuracy model. Regarding the effect of specific indices, average of all cosines explained most of the variance in both the RT and accuracy models. Number of cosines, slope, third highest cosine, and highest cosine indices were also successful predictors, but explained 1% or less of the variance in processing scores. In sum, the answer to the second research question is that the L2 indices were stronger predictors of lexical processing, especially Word2vec indices. Overall, words that are more semantically rich, that are less distinct or occur in fewer distinct contexts, and words that are acquired later are processed faster or more accurately.

3.5 Discussion

Meaning has been regarded as a major driving force in structuring the mental lexicon (Landauer & Dumais, 1997; Lund & Burgess, 1996). Related lexical items are conceivably stored together because we experience these items together or in similar contexts; that is, speakers are tuned to the distributional properties in the input and develop networks where related items and items that behave similarly are clustered in the mental lexicon. Evidence from computational models and behavioral tasks supports these claims. Distributional semantic models based on word co-occurrence from large corpora have been successful in modeling semantic relationships of words, which is taken as evidence that they may follow the same learning process as humans (Jones et al., 2012; Landauer & Dumais, 1997; Lund & Burgess,

1996; Mikolov, Chen, et al., 2013). Behavioral evidence from both L1 and L2 studies has also suggested that semantic variables have a significant impact on word processing (Bates et al., 2001; Berger, Crossley, & Skalicky; 2019; Cuetos & Barbón, 2006; de Groot et al., 2002; Hamrick & Pandža, 2020; Skalicky et al., in press). Despite the evidence, measures of lexical sophistication that embody semantic context are scarce, especially semantic context indices based on L2 corpora. To address these gaps, Study 2 of this dissertation tests corpus-based L2 semantic context indices derived from two computational approaches: Latent Semantic Analysis (Landauer, 2007) and Word to Vector (Mikolov, Chen, et al., 2013). To validate the L2 indices, they were used as explanatory variables of L2 writing and L2 lexical sophistication data. They were also compared to similar L1 indices to test their explanatory power beyond what L1 indices can explain.

The first validation step entailed the use of the L2 semantic indices and similar L1 indices as explanatory variables of writing quality as measured by holistic human ratings of essay quality. The independent and integrated TOEFL essays and their scores, which have been extensively adopted in L2 writing studies (e.g., Biber & Gray, 2013; Enright & Tyson, 2008; Friginal et al., 2014; Guo et al., 2013), were used as baselines for L2 writing quality models. This step answers the first research question, which asked to what extent the L2 and L1 semantic context indices predicted L2 writing proficiency. The models suggested that the L2 semantic indices were significantly predictive of L2 writing. They explained up to 18% of the independent essay scores and 10% of the integrated essay scores. The L2 indices were more predictive than the L1 indices, which did not explain any variance of the independent scores and explained only 5% of the integrated scores. The contribution of the L1 indices to the combined models were also low (i.e., approximately 1%).

A combination of different L2 semantic indices helped explain the independent and integrated essay scores, suggesting that each index represents a different aspect of semantic context. For the independent task, *average of all cosines* and *number of cosines above .3* derived from EF-CAMDAT were both significant predictors. For the integrated task, three indices based on EF-CAMDAT (i.e., *slope*, *number of cosines above .3*, and *highest cosine similarity*) and two indices based on TASA (i.e., *average top three* and *average of all cosines*) were significant predictors. The contribution of these indices is detailed below.

The index *average of all cosines* provided the strongest significant contribution to the independent model, explaining almost all the variance. *Average of all cosines* is a measure that synthesizes all relationships that a lemma has with other lemmas by averaging all cosine values of each related lemma to the target lemma between intermediate and mature models. This method allows for a developmental representation of semantic context; that is, this index accounts for semantic representations of earlier and later stages of learning as represented by the EF-CAMDAT proficiency levels. Lemmas with a high *average of all cosines* score tend to be semantically rich, occurring both in closed and unrestricted environments. Correlations with semantic variables, such as familiarity and concreteness, and frequency variables suggest that many of these lemmas are familiar, related to many different lemmas, and appear in several contexts. The writers who used less semantically rich lemmas scored higher in the independent essay. Appendix I provides an example of a high-scored and low-scored independent essay, with lemmas that contributed to higher scores highlighted in red. As in Study 1, the words highlighted in red are the lemmas below or above the mean scores of all test takers, depending on the relationship of the index with essay scores. The individual output for the index *average of all cosines* shows that both essays used relatively high semantically rich lemmas (i.e., scores were

on average 0.8 or higher) but the high-scored essay contained more lemmas that were less semantically rich, including "commonly," "dislike," and "feature."

The index *number of cosines above the .3 threshold* was also an explanatory variable of the independent scores. This index shows the number of close neighbors to the target lemma (i.e., lemmas that co-occur with the target lemma above a threshold). Because lemmas with a high *number of above .3 cosines* can be used in several distinct contexts, they represent items that are distinctly rich. For example, the lemma "corporate" used in the high-scored essay in Appendix I is closely related to 393 other lemmas, including "universalistic" (cosine = 0.895), "shareholder" (cosine = 0.888), and "divisive" (cosine = 0.886). The word "thing," which is less distinct, is only closely related to the lemma "refreshed" (cosine = 0.305). More proficient writers gave preference to more distinctly rich lemmas, as illustrated in Appendix I, which shows a clear concentration of distinctly rich lemmas in the high-scored essay, including "diversity," "corporate," "industry," and "supply."

There were no TASA indices that helped explain the independent essay scores either in the TASA independent or in the combined independent model. Also, the TASA indices had very low and non-significant correlations with the independent essay scores. The conclusion section below and Chapter 5 discuss potential reasons for the lack of effect of the TASA indices in the independent models.

In the integrated model, EF-CAMDAT *slope* was the strongest predictor of essay quality in both the EF-CAMDAT and combined models, explaining 6% of the variance in the combined model, which explained 10.4% of the scores. Other significant EF-CAMDAT indices included *number of cosines above .3* and *highest cosine similarity*. Lemmas that mature later (i.e., have higher *slope* scores) and are more distinctly rich (i.e., lemmas with a greater *number of cosines*

above .3) were related to higher integrated essay scores. However, the presence of lemmas that occurred in a highly distinct environment (i.e., they had a greater *highest cosine* score) was associated with lower essay quality. Two TASA indices contributed to the integrated models: *average of all cosines* and *average top three cosines*. The presence of more semantically rich lemmas, as measured by the index *average of all cosines,* and less distinct lemmas, as measured by the index *average top three cosines,* were associated with higher integrated essay quality. The effect of these indices is detailed below.

The effect of the index EF-CAMDAT *slope* in the integrated essays was in the expected direction. Lemmas with a higher slope mature later or appear in higher levels in the EF-CAMDAT corpus; therefore, they tend to be more sophisticated and specialized items that are expected to be used by proficient writers. Many of the lemmas with a high slope came from the source (see Appendix J for individual output for a high-scored and a low-scored integrated essay), but the more proficient writers added other sophisticated lemmas such as "completely," "certain," and "reasonable." The effect of the index EF-CAMDAT *number of cosines above .3* on the integrated scores was similar to its effect on the independent essays: more distinctly rich lemmas led to higher scores. The example in Appendix J shows that the high-scored essay contained lemmas that were more distinctly rich such as "beak" and "celestial," which, despite being from the source, were not used in the low-scored essay. Somewhat unexpectedly, the effect of the index EF-CAMDAT *highest cosine* on integrated scores was negative. A lemma with a close relationship with another lemma tends to be distinct, or more unique and specialized. However, they also tend to be more concrete, less ambiguous, and more imageable, which are characteristics of less sophisticated lemmas. Due to the low impact of this index, a strong trend was not observable in the individual output, but it seems that the use of less specialized lemmas

(i.e., lemmas with a low *highest cosine* score) that helps the writer reference and analyze the sources explains its negative impact on writing quality. Examples of these lemmas are found in the high-scored essay in Appendix J and include "completely," "speak," "theory," "lecture," and "lecturer." It is important to note that this index did not contribute to the combined model; that is, in the presence of other predictors, *highest cosine similarity* was irrelevant.

The effect of the TASA index *average of all cosines* in the integrated model was opposite to the effect of EF-CAMDAT *average of all cosines* in the independent model such that the use of more semantically rich lemmas (i.e., lemmas with a rich network of semantic relationships) led to higher scores in the integrated essay. The example in Appendix J shows that this was not always the case (i.e., high-scored essays sometimes contained fewer semantically rich words), but analyses of other examples suggested that it was the effective use of semantically rich lemmas from the source such as "bird," "mountains," "rivers," and "distances" that caused this positive relationship. The effect of the TASA index *average of top three cosines* was similar to the effect of the EF-CAMDAT index *highest cosine*, which are both top cosine measures related to semantic distinctness. Proficient writers used less distinct words which, based on individual output analyses, seem to help the writer compare and describe the sources. In Appendix J, the lemmas "specific," "tries," "speaks," and "fact," which are only present in the high-scored essay, corroborates this interpretation. The only TASA index that contributed to the combined model was *average of top three cosines*, explaining only 1% of the variance in the integrated scores.

A comparison with previous L2 writing studies is not entirely possible due to the limited number of studies that have used semantic and contextual distinctness indices as lexical sophistication benchmarks. However, a few considerations can be made. Crossley and McNamara (2012) and Berger et al. (2017) found that L2 users at a higher proficiency level used

more distinct words in writing and in speaking. This study confirms these findings and adds more to them. More proficient writers gave preference to lemmas that are more distinct (i.e., they have close relationships with other lemmas) and that are less semantically rich (i.e., lemmas with a more restricted network of semantic relationships), as measured by the *average of all cosines*, which had the highest impact in the independent essay models. These writers also opted for lemmas that develop semantic representations in later learning stages, as suggested by the *slope* index, which had the highest impact in the integrated essay models. Similar to Study 1, the impact of semantic context indices was higher in the independent essay. This was probably due to the confounding effect from integrated words in the integrated essays. This effect was particularly noticeable in this study, which only dealt with lemmas. Several of the distinct lemmas used in the essays came from the source affecting the impact of some indices, especially the weaker predictors such as *number of cosines above .3* and *highest cosine similarity*.

The results from the word processing models suggest a role of semantic context in L2 word processing. The EF-CAMDAT semantic context indices explained 8.6% of the variance in the speed of processing (i.e., how fast L2 users judged a word to be a pseudoword or a real word), and 8.5% of the variance in processing accuracy (i.e., how accurately L2 users judged a word to be a pseudoword or a real word). The TASA indices, on the other hand, explained less than 1% of word processing behavior. Lemmas that are more semantically rich, as measured by the EF-CAMDAT *average of all cosines*, are processed faster and more accurately. This index had the highest impact on both the reaction time and accuracy models, explaining 8.4% and 6.6% of the variance, respectively. Contrary to this effect, the index TASA *average of all cosines* suggested that less semantically rich lemmas are processed faster. The impact of this index was much lower (i.e., it explained 1.4% of the RT scores in the combined model), and it did not

affect accuracy. As illustrated in Appendix K, which features the 100 words that were processed faster and more accurately and the 100 words that were processed slower and less accurately, the effect of the EF-CAMDAT *average of cosines* is apparent: the more semantically rich words are concentrated among the lemmas with low RT and high accuracy scores. The effect of the index TASA *average of all cosines* is less clear, though, with semantically distinct lemmas among the lemmas that are processed faster and slower. Lemmas such as "bear," "foot," "snake," "hard," "city, and "book" which were indexed as semantically rich by EF-CAMDAT and were, therefore, processed more efficiently, were indexed as less semantically rich by TASA, suggesting that these lemmas may be not extensively represented in TASA for appropriate semantic representations to be developed.

Indices of semantic distinctness as measured by top cosines (i.e., *highest cosine similarity, third highest cosine similarity*, and *average top three cosine*), albeit weak, also surfaced as predictors of L2 lexical processing. As suggested by the EF-CAMDAT *third highest cosine* and TASA *average of top three cosine,* lemmas that are more distinct (i.e., occur in unique environments) have a processing disadvantage. However, the EF-CAMDAT *highest cosine similarity* index suggested that lemmas that occur in a highly distinct environment may be easier to process. It is possible that this effect was brought by lemmas that concurrently occur in a few distinct environments, as suggested by the index *highest cosine similarity,* and in less distinct environments. This seems to be the case with the lemmas "fireplace," "myth," "slack," and "snore," which had high distinctness scores. It is worth pointing out that *highest cosine* was one of the weakest semantic distinctness indices, explaining less than 1% of the variance in reaction time and accuracy scores.

The effect of EF-CAMDAT *slope* was also unexpected. Lemmas that are acquired later (i.e., they matured later in the semantic models) were produced more accurately. The effect of this variable was also small (i.e., it explained 0.2% of the variance in accuracy scores). Words that are acquired later and that were processed accurately included "nation," "cheese," "playground," "nose," "coin," and "list." These are seemingly common lemmas in an L1 environment, reflecting the experience of most participants in the lexical decision data, but they might have been featured in the *Englishtown* tasks at later levels. Lemmas similar to these might have caused this marginal, yet significant, positive effect of *slope* in the accuracy scores.

The index EF-CAMDAT *number of cosines* had an expected effect on accuracy scores. Lemmas that were less distinctly rich (i.e., they occurred in fewer distinct contexts), were produced more accurately. For example, lemmas that are not restricted to specific contexts such as "text," "secret," "response," and "similar" were processed more accurately. The effect of this variable was also marginal (i.e., it explained 0.2% of the accuracy scores).

The findings from the lexical decision models mostly mirror previous findings. Lexical processing research with L2 data has found that lemmas that are semantically rich are processed faster and more accurately (Berger, Crossley, & Skalicky, 2019; Hamrick & Pandža, 2020; Skalicky et al., in press). These semantically rich lexical items tend to appear in more semantic contexts and are related to more lemmas; therefore, they develop more entrenched representations and more connections with other items in the mental lexicon, which facilitate processing. The present study also found that, overall, more distinct lemmas, as measured by top cosine indices, have a processing disadvantage. Interestingly, even lemmas that occur in several distinct contexts (i.e., *number of cosines above .3* was high) were processed more slowly. This was the case for words like "muck," "triumph," and "gallop," which all had more than 4,000

related lemmas above the .3 threshold. In other words, these multiple contexts may not be enough to facilitate processing because these lemmas may still be limited to specific contexts. To the author's knowledge, no studies have investigated whether semantically rich but distinct lemmas such as the ones exemplified above have a processing disadvantage indeed. Future studies could explore these questions in factorial designs with the use of the indices introduced in this study.

3.6 Conclusion and Limitations

The results of Study 2 suggest that the semantic context indices introduced in this dissertation provide unique representations of semantic relationships, including representations of semantic richness, distinctness, and maturation, that can be successfully used in the study of L2 writing and lexical processing. The index *average of all cosines,* which incorporates semantic relationships from intermediate and mature models and all related lemmas, was particularly predictive. It was the strongest predictor in three major models (i.e., independent models, RT models, and accuracy models), and, when tested as the only predictor in the independent model, it explained 16.7% of the scores, a variance not explained uniquely by any lexical sophistication index tested in this dissertation. This may suggest that semantic context indices that include information about the development of lexical representations (i.e., cosine information from intermediate and mature models) and all the relationships that a lemma has with other lemmas (i.e., all cosines) provide a powerful representation of semantic knowledge.

A few considerations regarding the advantage of Word2vec and EF-CAMDAT indices should be noted. Both L2 Word2vec and LSA indices surfaced as predictors of both L2 writing and L2 word processing, but, similar to previous studies (Altszyler et al., 2018; Crossley, Kyle, et al., 2019), there was an overall preference for Word2vec indices. This might be due to the

local nature of Word2vec, which captures relationships from close lemmas. There is also evidence that Word2vec better represents human cognition when multiple topics are included, as is the case of EF-CAMDAT, whereas LSA performs better with domain-specific texts (Altszyler et al., 2018). The clear advantage for EF-CAMDAT indices over TASA indices can be due to several factors. First, TASA indices were based on LSA, which, as shown in this study, seems to perform worse than Word2vec regarding the development of lexical sophistication norms. Second, TASA represents the linguistic reading experience of average American students; therefore, it may not be useful to explain L2 lexical proficiency. Thirdly, TASA is based on edited texts such as textbooks, which are not the best representations of natural language (McDonald & Shillcock, 2001). Fourthly, EF-CAMDAT contained more lemma information than TASA, which might have given this index an advantage in the models. However, evidence from the combined RT and accuracy models, which only included the overlapping items, suggests that this advantage may not have been what caused the discrepancies in the effect of TASA and EF-CAMDAT. Lastly, TASA contains repeated samples of texts, which might have interfered with the semantic representations. Therefore, more semantic context indices of lexical sophistication based on different corpora need to be developed and tested to judge differential effects of L2 and L1 indices. Specifically, DMSs could be developed from corpora such as COCA Fiction and COCA Academic, which have been used extensively in L2 research as predictors of speaking and writing.

A few limitations should also be noted. Like any representations of lexical sophistication, LSA and Word2vec are highly dependent on the corpora that are used; therefore, the EF-CAMDAT indices tested here are restricted to written language from L2 users in an educational context (i.e., the online language learning platform *Englishtown*). Many of the L2 indices scores

reflect lemma relationships based on the *Englishtown* tasks. For example, the word "departure" co-occurs frequently with "lounge," "airport," and "stopover" possibly due to tasks requiring the use of these words. This influence may be what gave the L2 indices an advantage in analyzing the TOEFL essays, which are also task-based. Also, even though more than 30 million words were included in the development of the semantic spaces, dimensional semantic models have been shown to provide more accurate semantic representations with hundreds of millions of words (Mikolov, Chen, et al., 2013; Mikolov, Yih, et al., 2013). Other limitations related to the DSMs used here is that they do not account for word order and polysemous words (Landauer, 2007). Also, both LSA and Word2vec focus on highly frequent words, which limits the understanding of semantic representations of less frequent items (Jamieson et al., 2018). The elimination of non-standard forms through spellcheckers and dictionaries also eliminated neologisms and non-standard forms that can be useful for understanding L2 semantic relationships. Finally, the analysis of the integrated essays seemed to have been confounded by integrated words, especially for weaker predictors, generating contradictory findings. A better approach might have been to control for the integrated words to gauge the test-takers' lexical knowledge.

In addition to the limitations stated above, some specific considerations regarding the representativeness of some indices need to be stated. Even though this study treated the L2 semantic context indices as measures of distinctness or richness, most of them were based on limited semantic information that made it difficult to fit them into a single category. For example, top cosines indices such as *highest cosine* or *second highest cosine* provide information about the relationship between two lemmas while ignoring other relationships. As discussed above, lemmas that occur in distinct semantic environments may also be present in several

others, not being as distinct as top cosines scores might suggest. The contradictory findings for some of these variables confirm that some of these indices may be limited in their representation of semantic relationships, and that more holistic measures such as *average of all cosines* and *slope* might be more appropriate for indexing lexical sophistication. Other indices such as *average of all cosines* from mature models and *number of all cosines* should be tested as semantic context indices to verify whether more holistic representations are more representative of semantic context.

Despite the limitations, the L2 semantic context indices explained up to 18% of the L2 writing scores and up to 8.6% of the L2 lexical processing data. These findings suggest that semantic context information that resembles human knowledge of semantic relations can be successfully used in the automatic assessment of writing quality and word processing. It can also be used to test new hypotheses regarding processing, such as the role of the quality of semantic connections (i.e., lemmas with a network of closer or more distant relationships) in lexical processing.

4 STUDY 3: DEVELOPING AND TESTING L2 WORD RECOGNITION INDICES

Recognizing a word is one of the most fundamental processes of language comprehension (de Groot, 2011; Batia Laufer, 1992). Due to its importance in comprehension, word recognition has been one of the most investigated phenomena in psycholinguistics (e.g., Balota & Chumbley, 1985; Brysbaert et al., 2000; de Groot et al., 2002; Morrison & Ellis, 2000), driving the development of important theories regarding first language (Rumelhart & McClelland, 1982) and second language lexical processing (Dijkstra et al., 2019). Among the crucial contributions that studies based on word recognition have brought to the understanding of bilingualism are that L1 and L2 word processing are interconnected (Kerkhofs et al., 2006; Lagrou et al., 2011) and that degree of exposure, as opposed to an inherent lower capacity to learn a language in adulthood, explains differential effects in L1 and L2 processing (Monaghan et al., 2017).

The main unit of analysis of lexical processing studies has been single words. Words are of interest because they contain a limited set of constituents such as letters and phonemes that can be easily manipulated in research (Balota et al., 2012).[13] Behavioral ratings and responses to word stimuli have provided valuable information about lexical processing and the characteristics of words that facilitate processing (Assche et al., 2020; de Groot, 2011). These ratings and processing information have been particularly relevant in the field of natural language processing, whose main goal is to simulate human cognition through the use of natural language and behavioral data (Dikli, 2006). Of relevance is the use of subjective behavioral ratings related to the psychological properties of words to the automatic analysis of L2 texts. These analyses

[13] It is worth pointing out that, despite the fact that single words have been common stimuli in psycholinguistic research, phraseological studies suggest that words alone do not have meaning. In phraseology, phrases such as n-grams and phrase frames are considered the fundamental unit of language (Sinclair, 2008).

have included rating-based indices of age of acquisition (e.g., Crossley & McNamara, 2009), word concreteness (e.g., Crossley et al., 2015), word familiarity (e.g., Guo et al., 2013), word meaningfulness (e.g., Crossley & McNamara, 2012), word associative context (e.g., Berger et al., 2017), and word imageability (e.g., Kyle & Crossley, 2016), which have been successfully used to explain L2 lexical proficiency. Recently, word processing information from word recognition tasks such as reaction time and word accuracy data have also been used in the analysis of L2 language production, surfacing as significant predictors of lexical proficiency in L2 speaking and writing (Berger, Crossley, & Kyle, 2019; Kyle et al., 2018). An advantage of this method is that it uses respondents' objective online word processing information instead of subjective judgement or interpretation of stimuli as it is the case with rating-based indices.

Despite the contribution of the above-mentioned benchmarks in the understanding of L2 lexical production, the majority of these benchmarks have been based on L1 behavioral data (i.e., ratings of word properties and processing related to the respondents' first language), reflecting characteristics of monolingual processing. Even though L2 processing is not qualitatively dissimilar to L1 processing (i.e., the mechanisms are the same), word recognition studies have shown repeatedly that there are important quantitative differences related to the reduced exposure and unique circumstances under which a second language is learned (de Groot, 2011). Therefore, L2 behavioral data are needed to quantify these differences. To address the gap in the scarcity of robust L2 lexical processing data that directly represents L2 lexical processing, word recognition norms for about 5,000 words were collected from L2 users of English and tested as potential automatic indices of lexical sophistication. Specifically, reaction time and accuracy information from a word naming task performed by L2 users were compared to similar L1 and L2 word recognition norms and tested as explanatory variables of L2 writing quality. In doing

so, Study 3 addresses question number three of this dissertation regarding the validity of the L2 word recognition norms and their predictive power.

4.1 Visual Word Recognition and Lexical Processing

Lexical proficiency has been extensively investigated through online psycholinguistic tasks that require the production, recognition, association, and sorting of lexical items (Menn & Dronkers, 2017). These tasks allow researchers to investigate lexical processing from the initial stages of lexical access to the depth of lexical knowledge (i.e., the strength of lexical network connections) in L1 and L2 users (Leow et al., 2014). One of the most investigated aspects of lexical proficiency is word recognition. Also known as lexical access, word recognition refers to the match between the input word (i.e., oral or written) and its form in the mental lexicon, leading to the access of semantic, morpho-syntactic, and orthographic information about the word (de Groot, 2011). The visual word recognition paradigm, which involves the recognition of written words, has been particularly helpful and commonly used in the investigation of lexical access.

Word naming (i.e., a word reading task) and lexical decision tasks have been the most used visual word recognition tasks for the study of isolated word processing (Assche et al., 2020; Balota et al., 2012; de Groot, 2011). These tasks have been used in the investigation of the lexical variables that influence processing. By far, the most investigated variables are those related to linguistic experience such as word frequency, frequency of contexts (i.e., range), frequency of semantic contexts, and age of acquisition, which have been shown to have a strong impact in lexical access both in L1 and in L2 processing (Balota & Chumbley, 1985; Berger, Crossley, & Skalicky; 2019; Brysbaert et al., 2000; de Groot et al., 2002; Hamrick & Pandža, 2020; Morrison et al., 2002; Morrison & Ellis, 2000; Muncer et al., 2014; Shibahara et al., 2003;

Skalicky et al., in press), with a stronger frequency effect reported in the L2 (e.g., Brysbaert et al., 2017; Lemhöfer et al., 2008). Variables related to word characteristics such as frequency of sound combinations (Muncer et al., 2014), word length (Balota & Chumbley, 1985; Morrison & Ellis, 2000), number of morphemes (Muncer et al., 2014), and orthographic neighborhood (Morrison & Ellis, 2000; Muncer et al., 2014) are also important predictors of lexical processing in word naming, where production is required. These findings suggest that greater exposure to words, sounds, and morphemes results in faster naming and more accurate pronunciation. Semantic properties have also been investigated as predictors of lexical processing, including word concreteness (Richards, 1976; Skalicky et al., in press), imageability (Cortese & Schock, 2013; de Groot et al., 2002), meaningfulness (Colombo et al., 2006; Kristofferson, 1957), and familiarity (Colombo et al., 2006). Overall, these studies have suggested that more imageable, concrete, familiar, and meaningful words are processed faster.

Models of lexical processing have shown that the variables mentioned above affect both L1 and L2, indicating that L1 and L2 processing is qualitatively similar; however, processing scores suggest that L2 users tend to be slower and less accurate (de Groot et al., 2002; Kaur, 2017). Studies that have controlled for participants' language proficiency have suggested that the "disadvantage" seen in L2 users is best explained by linguistic experience, with more proficient L2 users being faster and more accurate (Brysbaert et al., 2017; Jared & Kroll, 2001; Lemhöfer et al., 2008). There is also evidence suggesting that the performance of experienced bilinguals approximates the performance of experienced monolinguals (Johns et al., 2016). Because L2 users can have less cumulative experience with the L2 than with the L1, lexical representations tend to be weaker, resulting in lower accuracy rates and higher reaction time scores (de Groot, 2011; Kaur, 2017).

The studies described above have helped answer important questions regarding connectionist theories, including the contribution of linguistic experience in lexical access (Chater & Christiansen, 1999). The effect of frequency-based variables in word recognition has suggested that items that are experienced more frequently and in more semantic contexts have stronger representations in the mental lexicon (Balota & Chumbley, 1985; Hamrick & Pandža, 2020; Morrison et al., 2002). The effect of morphological and phonetic variables have supported the hypothesis that activation of phonological and morphological representations occur during lexical access (Chater & Christiansen, 1999; Dijkstra & Heuven, 2002; Zhou et al., 2010). Studies that have used word recognition tasks with semantically related words (i.e., semantic priming tasks) have found a facilitation effect for processing semantically related words presented sequentially, confirming that related word candidates are activated in conjunction with the target word and are, therefore, connected in the mental lexicon (Perea & Gotor, 1997; Segui & Grainger, 1990). Studies that used word recognition tasks with interlingual homographs (i.e., words that have similar orthographic form, but different meanings across languages) have found competition effects, which suggests the cross-activation of languages in an integrated bilingual system (Dijkstra et al., 1998; Kerkhofs et al., 2006; Lemhöfer & Dijkstra, 2004). Despite much evidence, models of bilingual processing have been incomplete due to the lack of sufficient evidence (Dijkstra et al., 2019), requiring more investigations and the use of larger L2 datasets to test hypotheses raised by monolingual studies.

4.2 Psycholinguistic Word Information and L2 Writing

L2 studies of lexical proficiency have greatly benefitted from psycholinguistic word information derived from behavioral tasks. Assuming that natural language from timed writing and natural speech is influenced by the constraints imposed by lexical processing, the analysis of

L2 texts using word processing information as benchmarks has provided a gateway, albeit indirect, into the L2 mental lexicon. This method has opened up opportunities to analyze natural language from a processing perspective across proficiency levels. Behavioral ratings incorporated into L2 writing studies has included age of acquisition (e.g., Crossley & McNamara, 2009), word concreteness (e.g., Crossley et al., 2015), word familiarity (e.g., Guo et al., 2013), word meaningfulness (e.g., Crossley & McNamara, 2012), word associative context (e.g., Berger et al., 2017), and word imageability (e.g., Kyle & Crossley, 2016). These studies have suggested that less proficient writers give preference to words that are more concrete (Crossley et al., 2015), more imageable (Crossley, Kyle, et al., 2014), less polysemous (Kyle et al., 2018), more familiar (Crossley & McNamara, 2012), less specific (Kyle et al., 2018), and have more associations with other words (Crossley & McNamara, 2012). All these findings have indicated an L2 writing developmental path from less sophisticated to more sophisticated words.

Recent studies into the automatic assessment of L2 writing have benefited from word recognition information (i.e., reaction time and accuracy data from word recognition tasks) as predictors of lexical proficiency in L2 speaking (Berger, Crossley, & Kyle, 2019) and L2 writing (Kyle et al., 2018). The advantage of this method is that these measures are not based on the respondents' subjective judgement or interpretation of stimuli such as familiarity and age of acquisition judgements, but objective online processing. Kyle et al. (2018) found that these online word recognition benchmarks can help explain holistic scores of lexical proficiency of L2 writing. Words that were processed more efficiently by L1 users were associated with lower L2 lexical proficiency. However, the indices used so far have been based on L1 word recognition measures, which have been shown to be quantitatively different from L2 word recognition measures (Diependaele et al., 2013; Monaghan et al., 2017). This dissertation addresses this

limitation by testing word recognition indices developed from L2 behavioral data collected for Study 3.

4.3 Research Question

Study 3 was designed to answer the third research question of this dissertation regarding the validity of L2 recognition indices, which were compared to similar L2 and L1 indices, and the predictive power of L2 word recognition indices as explanatory variables of L2 writing scores by themselves and in comparison with similar L1 indices. The following specific research questions guided Study 3:

1) How do L2 word recognition indices compare to similar L1 and L2 word recognition indices derived from behavioral data?
2) To what degree do L2 and L1 word recognition indices derived from behavioral data predict L2 writing proficiency?

4.4 Methods

Study 3 develops and tests the validity and predictive power of L2 word recognition indices. For the development of the L2 indices, lexical processing data from L2 users using a word naming task was gathered for 4,998 words. The L2 users included in this study answered a background questionnaire, completed a lexical decision task aimed at assessing vocabulary proficiency, and performed a word naming task. The reaction time and accuracy data from the word naming task were used to develop word recognition indices that were tested in models of L2 writing proficiency and compared to similar word recognition datasets. In what follows, the participants, vocabulary test, word naming task, indices developed from the word naming data, L1 indices used for comparison purposes, and the outcome variables used in the statistical analysis are described.

4.4.1 Participants

The participants were students at Georgia State University (GSU) who used English as a second language. They were recruited through classroom visits, advertisements in social media groups connected to GSU, and flyers around campus. Undergraduate, graduate, and students enrolled in the Intensive English Program (IEP) at GSU were accepted to participate in the study. Only IEP students at the highest level (i.e., level 5 of a 5-level program) were recruited since the purpose was to gather lexical representations from proficient speakers. A total of 94 students, 56 females and 38 males, whose ages ranged from 18 to 76[14] (mean = 26.25), participated in the study. Participants reported having studied English from one year to 28 years (mean = 11.1 years), lived in the USA from one month to 23 years (mean = 3.92 years), and used English 30% to 100% of the time (mean = 70.5%) at the time of the data collection. The most representative first languages were Spanish (N = 20), Portuguese (N = 19), Chinese (N = 12) and Korean (N = 9). Participants came from 40 different countries and had 33 different first languages. The number of participants per country is provided in Table 4.1 below.

Table 4.1 Distribution of Participants per Country

Country	***Number of Participants***	***Country***	***Number of Participants***
Brazil	17	Ecuador	1
China	14	Eritrea	1
South Korea	9	Georgia	1
Colombia	5	Germany	1
Mexico	4	Ghana	1
Hong Kong	3	Greece	1
Iran	3	Haiti	1
Saudi Arabia	3	Indonesia	1
Chile	2	Japan	1
France	2	Latvia	1
Ivory Coast	2	Lebanon	1
Nigeria	2	Madagascar	1

[14] The 76 year-old participant was an outlier.

Venezuela	2	Mongolia	1
Vietnam	2	Nepal	1
Angola	1	Pakistan	1
Bangladesh	1	Peru	1
Benin	1	Spain	1
Cuba	1	Thailand	1
Curacao	1	Tunisia	1
Dominican Republic	1	Turkey	1

4.4.2 Vocabulary Proficiency

Beyond their placement in the top level of an IEP program and fulfillment of the university language requirements, participants' vocabulary proficiency was also measured through a lexical decision task (i.e., a word-non-word decision task) adapted from Lemhöfer and Broersma (2012) called LexTALE (i.e., Lexical Test for Advanced Learners of English). This test was selected for the acceptable to high reliability reported in Lemhöfer and Broersma (.81 for Dutch participants and .67 for Korean participants) and for being a short proficiency test that would not significantly extend the time participants spent in the lab. The adapted task consisted of 6 practice trials, followed by 30 words (e.g., "scornful") and 30 non-words (e.g., "mensible"). It had 10 more non-words and 10 fewer words than the original. All non-words obeyed phonotactic and orthographic English rules. The adapted LexTALE was presented using E-prime (i.e., a software designed for behavioral research). Participants would see a string of characters and then were given a maximum of 10 seconds to judge whether the stimulus was a real word or not. They did so by pressing a green button if they considered the character string to be a word and a red button if they considered the character string to be a non-word using a Serial Response (SR) box. The SR box from Psychology Software Tools Inc. (www.pstnet.com/products/SRBOX/default.htm) is a device designed for experiments that require precise RT calculation not afforded by computer keyboards. A desktop PC in the

psycholinguistic lab at the Department of Applied Linguistics at GSU (i.e., L-PAL lab) was used for the task. Two measures of vocabulary proficiency were derived from the adapted LexTALE: reaction time and accuracy. Students' accuracy ranged from 51% to 92% (mean = 71%). Reaction time ranged from 762ms to 4458ms (mean = 2057.8ms). The Cronbach's alpha for this task was .70, which is considered acceptable.

The participants were randomly split into three groups to read three different sets of words from the word naming task. To check whether they were at the same proficiency level and that word reading information was comparable, the three groups of participants were compared using the LexTALE accuracy and reaction time. As illustrated in Table 4.2 below, the three groups had the same level of lexical proficiency, as suggested by the non-significant results of independent samples t-tests.

Table 4.2 Comparisons Between the Three Groups of Participants

	LexTALE RT List 2	***LexTALE RT List 3***	***LexTALE Accuracy List 2***	***LexTALE Accuracy List 3***
LexTALE RT List 1	t = –0.54737, df = 60.616, p = 0.5861	t = 0.29456, df = 58.91, p = 0.7694	–	–
LexTALE RT List 2		t = 0.82224, df = 59.901, p = 0.4142	–	–
LexTALE Accuracy List 1	–	–	t = –1.1135, df = 58.101, p = 0.2701	t = –0.24743, df = 57.155, p = 0.8055
LexTALE Accuracy List 2	–	–		t = 0.8193, df = 59.869, p = 0.4159

4.4.3 Word Naming Task

A word naming task, or a word reading task, is a psycholinguistic experiment used to measure early stages of lexical processing (Balota et al., 2007; Ferrand et al., 2011). It consists of single words presented on a screen that are read aloud by subjects under some time pressure. For this dissertation, the word naming task was adapted from the English Lexicon Project, a multi-university project developed by Balota et al. (2017), who gathered word processing information from more than 800 participants. The word naming task in the ELP involved 400 participants whose L1 was English. Each participant named approximately 2,500 words, resulting in a data set with reaction time and accuracy information for 40,481 words. For this study, 4,998 words were selected, and the procedures were adapted to the L2 population, as described below.

4.4.3.1 Word Selection

The words selected for the word naming task were primarily based on the words used in the lexical decision task developed by Berger, Crossley, and Skalicky (2019), who randomly selected 3,318 content words from the ELP project (Balota et al., 2007). Since only a partial replication of the ELP was feasible for this dissertation, the 3,318 words used by Berger, Crossley, and Skalicky (2019), and 1,680 additional words from the ELP were used. The additional 1,680 words from the ELP were selected based on high word frequency to increase the likelihood that the L2 participants were familiar with them. Only words in the top 10,000 words from the COCA Spoken corpus were considered for selection. Similar to Berger, Crossley, & Skalicky (2019), content words were prioritized as function words convey little meaning and have little predictive power (McDonald & Shillcock, 2001). Finally, proper nouns were also discarded for being words that many L2 users might not have encountered. Once these criteria were applied, the remaining words were selected to include a range of characteristics, including

semantic context and word properties such as word length. Table 4.3 below lists the indices that were considered in the selection, along with descriptive statistics for the 4,998 words. The EF-CAMDAT – raw frequency index was included for reference purposes.

Table 4.3 Lexical Characteristics of Word Naming Words

Indices	***Average***	***Medium***	***SD***	***Min***	***Max***
Length[a]	6.25	6.00	2.05	2.00	16.00
Orthographic Neighbors[a]	3.35	1.00	4.48	0.00	25.00
Phonological Neighbors[a]	7.30	3.00	9.36	0.00	48.00
Context Distinctiveness[b]	1.28	1.14	0.70	0.05	4.19
TASA SLA – Average of all cosines[c]	0.20	0.17	0.13	0.00	0.85
Polysemy[d]	7.25	5.00	6.88	1.00	75.00
MRC Concreteness[e]	450.97	447.00	115.97	190.00	670.00
MRC Familiarity[e]	541.18	545.00	48.10	228.00	657.00
MRC Imageability[e]	475.04	481.00	99.01	204.00	667.00
MRC Meaningfulness[e]	449.14	448.00	55.73	215.00	617.00
EF-CAMDAT – Raw frequency[f]	2230.55	294.50	9533.10	2.00	327743.00

[a] Indices from Balota et al's (2007) ELP database, [b] From McDonald and Shillcock (2001),
[c] From TAALES (Kyle et al., 2018), [d]From Fellbaum (1998), [e] From MRC database (Coltheart, 1981),
[f] Words with raw frequency below 55 were not in the top 10000 EF-CAMDAT list

The 4,998 words were distributed into three sublists, forming three lists with 1,666 words each. The 4,998 words were manually split by initial sounds and distributed across the three lists (e.g., words that started with the /f/ sound were equally distributed across lists) and across 7 sessions with 238 words each in each of the three lists. This procedure was meant to minimize the effects of morpheme stem practice (Balota et al., 2007). In each of the seven sessions, the words were presented in random order by E-prime. For half of the participants, the seven trials were reversed. A total of 32 observations per word were collected for list one, 32 observations per word for list two, and 30 observations per word for list three.

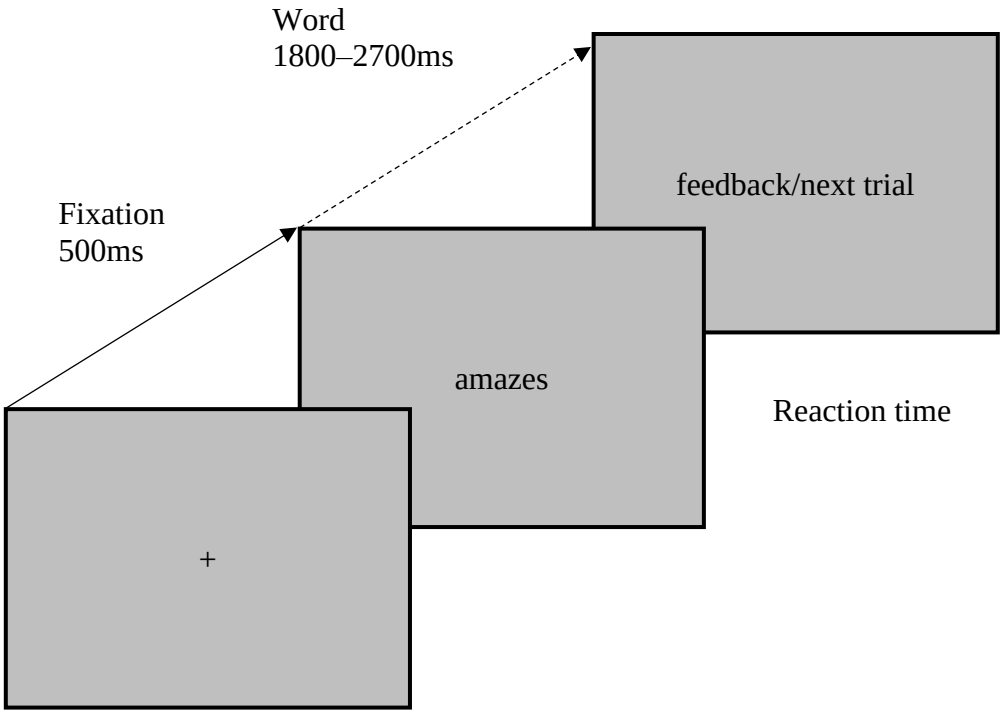
4.4.3.2 Task Procedures

The word naming task procedures were adapted from Balota et al. (2007) for the population of L2 users of English sampled in this dissertation. To account for the processing

difficulties found among L2 users when performing word naming tasks (Brysbaert et al., 2017; de Groot et al., 2002), the number of words was reduced from 2,500 to 1,666, and other minor adaptations were also performed as described below.

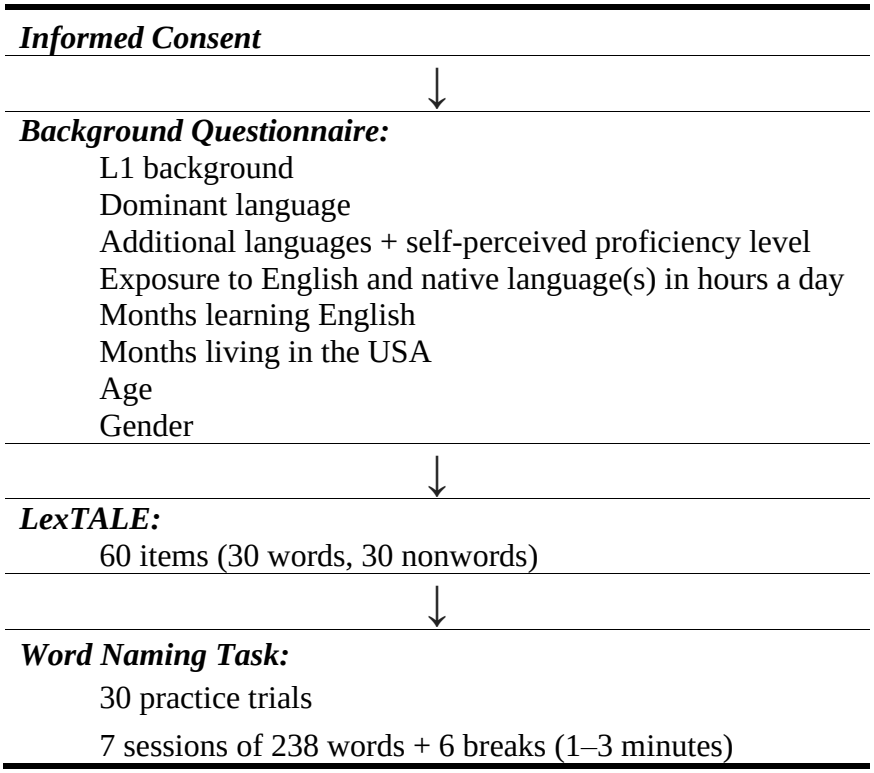
Participants started the word naming task with a practice session containing 30 words, which was followed by 7 sessions, with breaks between sessions that lasted 1–3 minutes. At the onset of the task, the participants were instructed to read the words as quickly and as accurately as possible. Each trial began with a fixation point in black presented in the center of a gray screen for 500ms; the word followed in a similar gray screen and remained on the screen until the next trial began. Because the SR box does not recognize when word reading stops, E-prime was programmed to keep each word on the screen from 1,800ms to 2,700ms, depending on the length of the word. Words with five or fewer characters remained on the screen for 1,800ms, words with 6–7 characters remained on the screen for 2,000ms, words with 8–9 characters remained on the screen for 2,300ms, and words with 10 or more characters remained on the screen for 2,700ms. Participants proceeded to read the word aloud. If reaction time was greater than 1,000ms, the participants received the following message: "Please read the words FASTER," which remained on the screen for 1,000ms. During the breaks, the participants saw a countdown. When the time was over, they were required to press the space key to start the next session. The word naming task lasted approximately one hour and 20 minutes. Note that in Balota et al. (2007), there were 40 practice trials, asterisks remained on the screen for 250ms, which were followed by a 50-ms tone and a 250ms dark interval, and the word remained on the screen for 250ms after word onset. Figure 4.1 illustrates a word naming task trial.

Figure 4.1 Schematic Illustration of a Word Naming Trial.



A desktop PC in the psycholinguistic lab at the Department of Applied Linguistics at GSU (i.e., L-PAL lab) was used to collect data from the task. A Shure PGA81-XLR Cardioid condenser microphone connected to a pre-amplifier and the PC was used to record the participants' pronunciation of words using the built-in recording function in E-prime, which generated a single audio file per stimulus word. A WH20XLR head-worn dynamic microphone connected to a pre-amplifier and an SR box was used to capture the onset of the word pronunciation for reaction time calculation. The consent form signature, background questionnaire, LexTALE, and the word naming task were all gathered on the same day, as illustrated by Figure 4.2, which shows the data collection procedures.

Figure 4.2 Data Collection Procedures



4.4.3.3 Word Naming Variables

Two variables were derived from the word naming task: reaction time and naming accuracy. Both are detailed below.

4.4.3.3.1 Word Naming Reaction time

Reaction time was operationalized as the time elapsed from the moment each word appeared on the screen to the onset of pronunciation. It was measured in milliseconds and automatically computed by the SR box. To minimize the effect of miscalculations of RT due to uncontrollable noises (e.g., noises from sneezing or yawning), or delayed pronunciations due to distractions, the data were cleaned following a few criteria. A minimum cut-off point for RT inclusion was set at 250ms, 50ms more than in Balota et al. (2007). This cut-off choice was based on the performance of the L2 participants (average of RT scores = 676.14ms, average of

RT SD scores = 200.37ms), whose reaction time was 46.3ms slower than the ELP participants for the same 4,998 words (ELP average of RT scores = 629.83ms, ELP average of RT SD scores = 132.22ms). Also, all RT values that were three standard deviations above or below the mean for their respective word were eliminated.

After the data were cleaned, the number of observations was reduced from 156,604 to 142,810 observations[15]. There was a minimum of 18 RT values per word and a maximum of 32; however, only two words had 18 values, and one word had 19 values. Mean RT values increased and standard deviations lowered after the data were cleaned (average of RT scores = 689.72ms, Min = 496.7ms, Max = 1108.2ms, and average RT SD scores = 183.44ms, Min = 74.55ms, Max = 449.25ms).

4.4.3.3.2 Word Naming Accuracy

Accuracy was operationalized as the participants' ability to accurately produce the stimulus word as judged by two human raters. In psycholinguistic studies, accuracy is defined as the "ability to accurately identify single words from print" (Pasquarella et al., 2015, p. 2). In other words, psycholinguistics is concerned with the underlying mental representation of the word, which can be manifested with native-like and non-native-like pronunciation. In psycholinguistic studies, accuracy is judged by checking whether the pronunciation matched the word (e.g., Balota et al., 2007; Shibahara et al., 2003). Like these studies, accuracy was based on raters' judgments on whether the audio file from the L2 learners matched the stimuli.

Two accuracy ratings for each word were performed. The author of this dissertation performed one rating, and the second rating was performed by workers on Mechanical Turk

[15] Most of the observations that were eliminated were 0 values that resulted from the SR box not calculating RT or participants skipping the words.

(Mturk), a crowdsourcing website that allows for affordable and reliable data collection (e.g., Schnoebelen & Kuperman, 2010; Sprouse, 2011). The Mturk task required workers to listen to the audio files (i.e., one audio file per word) and answer the following question: "Did the speaker say WORD?" where WORD is the word from the naming task. Each worker had to judge 50 words and earned $.40 for this judgment. To minimize issues with Mturk, ratings were only kept when a worker had an agreement rate of 80% or above with the researcher's judgments. Agreement occurred 91.58% of the time. Only the words with 100% agreement were kept. After this criterion was applied, 136,780 observations out of 156,604 remained[16]. The majority of the words (N = 4978) had 20 to 32 observations for accuracy. The remaining 20 words had 12 (N = 1), 15 (N = 3), 17 (N = 3), 18 (N = 8), and 19 (N = 5) observations. The average of word accuracy for the 4,998 words was 0.94, ranging from 0.05 to 1.00. The ELP average of word accuracy for the same 4,998 words was .99, ranging from 0.55 and 1.

4.4.4 L2 Word Recognition Indices

Reaction time and accuracy from the naming task were used to derive L2 norms of word naming (i.e., word recognition norms). Like Balota et al. (2007), four measures of word naming were developed, three based on RT scores (i.e., word naming response time, z-scored word naming response time, and word naming response time standard deviation), and one based on accuracy (i.e., word naming response accuracy). The word recognition indices developed for this dissertation are detailed below.

[16] Part of the observations were eliminated because of unintelligible pronunciations due to mumbling or whispering. A few audio files for the words were also empty because participants skipped the word.

4.4.4.1 L2 Word Naming Response Time

Word naming response time was the average RT across participants for each word, measured in milliseconds. An index with response time as z-score (i.e., the standardized average RT across participants for each word) and an index with the standard deviation (i.e., the SD calculated from the RT scores for each word) were also computed. High RT and SD scores are indicative of word processing difficulties. For example, the word "trousers," which has an RT mean score of 1013.9ms and SD mean score of 437.8ms is likely less cognitively entrenched and more difficult to access for L2 users than the word "computer," with an RT mean score of 530.6ms and SD mean score of 106.7ms.

4.4.4.2 L2 Word Naming Response Accuracy

Word naming response accuracy is the average naming accuracy for each word based on the judgment of two raters. Word naming accuracy also includes properties of online lexical processing. Words with higher accuracy are easier to process. For example, the word "courageous" had an accuracy of .52 (i.e., 52% of the respondents accurately recognized the word), would be considered more difficult than the word "situation," with an accuracy of 0.97. While RT can be a measure of how entrenched words are (i.e., the depth of lexical knowledge), with higher RT signaling less entrenchment, accuracy is more closely related to word knowledge (i.e., the breadth of lexical knowledge). More than one-third of the words had perfect accuracy scores (N = 1970), and most of the words (N = 4050) had an accuracy of 90% or above. This suggests that most of the words were known by the participants.

4.4.5 L1 Word Recognition Indices

The L1 indices adopted in this dissertation were calculated by TAALES (Kyle et al., 2018). The indices available in TAALES were derived from the English Lexicon Project

described above. In this dissertation, the ELP word information for content words, as opposed to all words, was used to match the content of the L2 indices. The ELP average of word RT for the 40,481 words used in TAALES is 722.82ms (Min = 507.8ms, Max = 2616ms), the ELP average of word SD is 178.54ms (Min = 25.2ms, Max = 639.4ms), and the ELP average of word accuracy is .93 (Min = .11, Max = 1). The L2 and respective L1 (ELP) indices used in this dissertation are listed in Table 4.4.

Table 4.4 L1 and L2 Word Recognition Indices

L2 Word Recognition Norms	***L1 (ELP) Word Recognition Norms***
L2 Word Naming RT	ELP Word Naming RT CW
L2 Word Naming RT (z-score)	ELP Word Naming RT (z-score) CW
L2 Word Naming RT (standard deviation)	ELP Word Naming RT (standard deviation) CW
L2 Word Naming Accuracy	ELP Word Naming Accuracy CW

4.4.6 Outcome Variables

The validation of the L2 word recognition indices was similar to the first validation step in Studies 1 and 2. L2 and L1 word recognition indices were computed for the TOEFL essays from the independent (N = 480) and integrated task (N = 480), and linear mixed-effects models were calculated using the TOEFL scores as the outcome variable and language as a random effect. For more information about the TOEFL essays, refer to Chapter 2.

4.4.7 Statistical Analysis

Correlations between the L2 norms and similar word recognition indices were performed to test convergent validity (i.e., the degree to which related measures are correlated). The ELP indices from Balota et al. (2007) and L2 lexical decision norms from Berger, Crossley, and Skalicky (2019) used in Study 1 and 2, were used as the benchmarks from which to judge the L2 word naming data. The correlation coefficients were used to address research question number one regarding the relationship between the L2 norms and related databases.

For the development of the L2 writing proficiency models, L1 (i.e., ELP) and L2 word recognition indices were computed for the words in the independent (N = 480) and integrated essays (N = 480). TAALES (Kyle et al., 2018) was used to compute the corresponding L1 word recognition indices. The L2 and L1 word scores were averaged, forming an average score for each index and each essay. Linear mixed-effects models were calculated using the integrated and independent TOEFL scores as the outcome variables, language as a random effect, and the word recognition index average scores as fixed effects. One independent and one integrated model for the L2 and L1 indices were developed, as well as a combined independent and a combined integrated model. The models were statistically compared using the r-squared difference test. Similar to Study 2, the control variables (i.e., age, gender, and topic) were not included because they have been shown to be non-significant predictors of TOEFL scores when tested against the unconditional model in Study 1.

As in Study 1 and Study 2, the forward approach to model development was adopted. A basic model with random effects (i.e., the unconditional model) was built, and predictors were added individually. Predictors were eliminated if model comparison statistics (i.e., likelihood ratio tests) showed that they did not improve the fit of the models. Model comparisons are in Appendix L, and the model with the best fit is reported in the results section below. Marginal and conditional r-squared for the fixed-effects model (i.e., the part of the LME model with fixed effects) and random-effects model (i.e., the part of the LME model with random effects) are reported, along with semi-partial r-squared for each fixed effect. The effect of the indices as reported by the semi-partial r-squared and the model comparisons were used as measures of how predictable the indices were and whether there were models that were significantly more predictable than others, which provides the answer to research question one related to the degree

of variance in writing scores explained by the L2 and L1 indices.

4.5 Results

This results section is divided into two main parts: database comparisons and L2 writing models, which address each research question.

4.5.1 Database Comparisons

To test whether the L2 indices were measuring similar lexical constructs as the L1 indices, correlations between the ELP and the L2 word naming indices were calculated. As shown in Table 4.5, correlations were medium for the RT scores and low for the SD and accuracy scores. The medium correlations for the RT scores suggest that the L2 RT indices were related to the L1 indices but included additional word processing information. The low correlation between SD scores was expected given that L2 users have a wide range of linguistic experiences that do not resemble L1 linguistic experiences, which are likely more homogeneous. Thus, L2 users are more likely to show greater variance in responses than L1 users. The low correlation for accuracy, which is more closely related to the breadth of lexical knowledge, was also expected to be more dissimilar to L1 accuracy data because L2 users have less cumulative experience with the L2 language and a different breadth of lexical knowledge.

Table 4.5 Correlations between the L2 Word Naming and ELP

	ELP Word Naming RT	***ELP Word Naming RT (z-score)***	***ELP Word Naming RT (SD)***	***ELP Word Naming Accuracy***
L2 Word Naming RT	0.42***	–	–	–
L2 Word Naming RT (z-score)	–	0.41***	–	–
L2 Word Naming SD	–	–	0.21***	–
L2 Word Naming Accuracy	–	–	–	0.28***

*** $p < .0005$

Correlations were also run between the L2 word naming indices and the L2 lexical decision indices available in Berger, Crossley, and Skalicky (2019) to test whether the indices

were measuring similar word processing information. As shown in Table 4.6 below, the correlation coefficients were low to medium, suggesting that there is some overlap between these two tasks, especially regarding RT. The low correlation coefficient between SD scores suggest that the L2 population in Berger et al. and this dissertation had different experiences with English (e.g., different proficiency levels and age of arrival). The low correlation regarding accuracy was also expected, given the differences between the tasks used in each study. In a word naming task, an accurate phonetic representation of the word is required for production to occur accurately, whereas, in a lexical decision task, only orthographic recognition is required.

Table 4.6 Correlations between the L2 Word Naming indices and the L2 Lexical Decision Indices

	Lexical Decision – RT Mean	***Lexical Decision RT (z-score)***	***Lexical Decision – RT (SD)***	***Lexical Decision – Accuracy***
L2 Word Naming RT	0.28***	–	–	–
L2 Word Naming RT (z-score)	–	0.37***	–	–
L2 Word Naming RT SD	–	–	0.09*	–
L2 Word Naming Accuracy	–	–	–	0.25***

*** $p < .0005$, * $p < 0.05$

These findings provide the answer to research question number one, which asked how the L2 indices relate to similar word recognition databases. The correlations showed that the RT indices, which are the primary indices of lexical access, are moderately correlated. This is indicative that there is some overlap in L1 and L2 processing, and that the L2 word naming and L2 lexical decision tasks seem to tap into similar word recognition processes. However, correlations with standard deviation scores were low, suggesting that the population in the three databases had very different linguistic experiences. These results were expected between the L1 population in ELP and the L2 population in this study; studies have repeatedly shown that linguistic experiences are different for L1 and L2 users. Also, the population in Berger, Crossley,

and Skalicky (2019) came from an online pool of L2 users with a wide range of linguistic experiences, differing from the university-student population in this dissertation. Finally, correlations with accuracy were also low, suggesting that differences in the breadth of lexical knowledge among the populations are also high. This finding is also expected given the dissimilarities in the participants' pool, as discussed above.

4.5.2 Writing Quality Models

This section is divided into four sub-sections: L2 writing proficiency models (i.e., models with the L2 word recognition norms as predictor variables), L1 writing proficiency models (i.e., models with the L1 word recognition norms), combined writing proficiency models (i.e., models with the L2 and L1 word recognition norms), and model comparisons. The writing proficiency model sections include correlations between the outcome variables and the indices, a model for the independent essays (i.e., the independent models), and a model for the integrated essays (i.e., the integrated models)

4.5.2.1 L2 Writing Proficiency Models

The L2 word recognition index scores for the TOEFL essays were checked for multicollinearity with a threshold set at $r \geq .7$. The L2 word naming RT z-score and word naming SD were eliminated because they were highly correlated with RT scores. Table 4.7 below shows the correlation scores between the writing tasks and the selected indices.

Table 4.7 Correlation Scores between Essay Scores and Selected L2 Word Recognition Indices

L2 Word Recognition Indices	***Independent***	***Integrated***
L2 Word Naming RT	0.381^{***}	0.176^{***}
L2 Word Naming Accuracy	$–0.273^{***}$	$–0.215^{***}$

*** $p < .005$

4.5.2.1.1 L2 Independent Model

The L2 independent essay model shows the effect of the L2 word recognition indices on the independent essay scores. Language was used as a random effect and the indices as fixed effects. Table 4.8 below shows the L2 independent model with the best fit along with the r-squared values and 95% confidence intervals for each fixed effect.

Table 4.8 L2 Independent Model with Best Fit

Random Effects	***Variance***	***SD***					
Language (intercept)	0.092	0.303					
Residual	0.589	0.768					
Fixed Effects	***Estimates***	***SE***	***t-value***	***p***	***R^2***	***95% CI***	
(Intercept)	–18.191	2.418	–7.524	0.05			
L2 Word Naming RT	0.034	0.004	8.992	<.005	0.13	0.19	0.08

The only fixed effect (i.e., word naming reaction time) that remained in the model explained 13% (marginal R^2= 0.13) of the variance, and the random effect explained 25% of the variance in independent scores (conditional R^2 = 0.25).

4.5.2.1.2 L2 Integrated Model

The L2 integrated essay model shows the effect of the L2 word recognition indices on the integrated essay scores. Language was used as a random effect and the indices as fixed effects. Table 4.9 below shows the L2 integrated model with the best fit and its statistics.

Table 4.9 L2 Integrated Model with Best Fit

Random Effects	***Variance***	***SD***					
Language (intercept)	0.118	0.344					
Residual	1.359	1.166					
Fixed Effects	***Estimates***	***SE***	***t-value***	***p***	***R^2***	***95% CI***	
(Intercept)	24.490	9.965	2.458	0.01	0.04	0.09	0.02
L2 Word Naming RT	0.010	0.005	2.109	0.03	0.01	0.03	0.00
L2 Word Naming Accuracy	–28.391	8.753	3.244	<.005	0.02	0.05	0.00

The fixed effects of the L2 integrated model explained 4% (marginal $R^2 = 0.043$) of the variance, and the random effect explained 12% of the variance in integrated scores (conditional $R^2 = 0.12$). Both word naming RT and accuracy were significant predictors, but accuracy scores had a higher impact according to the semi-partial r-squared value.

4.5.2.2 L1 Writing Proficiency Models

All L1 word recognition index scores for the TOEFL essays were checked for multicollinearity with a threshold set at $r \geq .7$. Table 4.10 below shows the correlation scores between the writing tasks and the selected indices.

Table 4.10 Correlation Scores between the Dependent Variables and the Selected L1 Word Recognition Indices

ELP Word Recognition Indices	***Independent***	***Integrated***
ELP Word Naming RT CW	0.354^{***}	0.241^{***}
ELP Word Naming SD CW	0.258^{***}	–
ELP Word Naming Accuracy CW	-0.110^{*}	0.031

*** $p < .0005$, ** $p < .005$, * $p < 0.05$, $p > .05$

4.5.2.2.1 L1 Independent Model

The L1 independent essay model shows the effect of the L1 word recognition indices on the integrated essay scores. Model statistics are reported in Table 4.11.

Table 4.11 L1 Independent Model with Best Fit

Random Effects	***Variance***	***SD***					
Language (intercept)	0.1074	0.3277					
Residual	0.6052	0.7779					
Fixed Effects	***Estimates***	***SE***	***t-value***	***p***	***R^2***	***95% CI***	
(Intercept)	–19.439	2.891	–6.723	<.005			
ELP Word Naming RT CW	0.037	0.005	7.951	<.005	0.11	0.16	0.06

The only fixed effect that improved the fit of the L1 independent model (i.e., word naming RT) explained 11% (marginal $R^2 = 0.11$) of the variance in independent scores, and the random effect explained 24% of the variance in scores (conditional $R^2 = 0.24$).

4.5.2.2.2 L1 Integrated Model

The L1 integrated essay model shows the effect of the L1 word recognition indices on the integrated essay scores. Model statistics are reported in Table 4.12.

Table 4.12 L1 Integrated Model with Best Fit

Random Effects	***Variance***	***SD***					
Language (intercept)	0.1045	0.3233					
Residual	1.3417	1.1583					
Fixed Effects	***Estimates***	***SE***	***t-value***	***p***	***R^2***	***95% CI***	
(Intercept)	–86.759	25.082	–3.459	<.005	0.06	0.11	0.03
ELP Word Naming RT CW	0.039	0.007	5.479	<.005	0.06	0.10	0.02
ELP Word Naming Accuracy CW	66.084	23.098	2.861	<.005	0.02	0.05	0.00

The fixed effects of the L1 integrated model explained 6% (marginal $R^2 = 0.06$) of the variance in essay scores, and L1 background explained 13% of the variance in integrated scores (conditional $R^2 = 0.13$). Two indices were significant predictors: word naming RT and accuracy. Semi-partial r-squared values suggested that RT had the highest impact in the model.

4.5.2.3 Combined Writing Quality Models

The L2 and L1 word recognition index scores for the TOEFL essays were checked for multicollinearity with a threshold set at $r \geq .7$. Table 4.13 below shows the correlation scores between the writing tasks and the selected indices.

Table 4.13 Correlation Scores between the Dependent Variables and the Selected L2 and L1 Word Recognition Indices

Word Recognition Indices	***Independent***	***Integrated***
L2 Word Naming RT	0.381^{***}	0.176^{***}
ELP Word Naming RT CW	0.354^{***}	0.241^{***}
L2 Word Naming Accuracy	$–0.273^{***}$	$–0.215^{***}$
ELP Word Naming SD CW	0.258^{***}	–
ELP Word Naming Accuracy CW	$–0.110^{*}$	0.031

*** $p < .0005$, ** $p < .005$, * $p < 0.05$, $p > .05$

4.5.2.3.1 Combined Independent Model

The combined independent essay model shows the effect of the L2 and L1 word recognition indices on the independent essay scores. Model statistics are reported in Table 4.14.

Table 4.14 Combined Independent Model with Best Fit

Random Effects	***Variance***	***SD***					
Language (intercept)	0.092	0.303					
Residual	0.580	0.761					
Fixed Effects	***Estimates***	***SE***	***t-value***	***p***	***R^2***	***95% CI***	
(Intercept)	–23.105	2.919	–7.916	<.005	0.15	0.20	0.10
L2 Word Naming RT	0.024	0.005	4.95	<.005	0.04	0.09	0.02
ELP Word Naming RT CW	0.018	0.006	2.956	<.005	0.02	0.05	0.00

The fixed effects in the combined independent model explained 15% (marginal R^2 = 0.15) and the random effect explained 26% of the variance in independent scores (conditional R^2 = 0.26). Both the L2 word naming RT and ELP word naming RT indices made a significant contribution to the model, but the L2 RT index had a stronger impact according to the semi-partial r-squared value.

4.5.2.3.2 Combined Integrated Model

The combined integrated essay model shows the effect of the L2 and L1 word recognition indices on the integrated essay scores. Model statistics are reported in Table 4.15.

Table 4.15 Combined Integrated Model with Best Fit

Random Effects	***Variance***	***SD***					
Language (intercept)	0.095	0.308					
Residual	1.330	1.153					
Fixed Effects	***Estimates***	***SE***	***t-value***	***p***	***R^2***	***95% CI***	
(Intercept)	–53.055	28.359	–1.871	0.06	0.07	0.13	0.04
ELP Word Naming RT CW	0.033	0.007	4.595	<.005	0.04	0.07	0.01
L2 Word Naming Accuracy	–21.954	8.686	–2.528	0.01	0.01	0.04	0.00
ELP Word Naming Accuracy CW	57.759	23.222	2.487	0.01	0.01	0.04	0.00

The fixed effects in the combined integrated model explained 7% (marginal R^2 = 0.07) of the variance in integrated scores, and the L1 background explained 13% of the variance (conditional R^2 = 0.134). Two L1 (i.e., word naming RT and accuracy) and one L2 index (i.e., word naming accuracy) entered the model.

4.5.2.4 Model Comparisons and Research Questions

The L2, L1, and combined models were statistically compared using the r-squared difference test. Table 4.16 shows the comparisons with the L2 independent model, and Table 4.17 shows the comparisons with the L2 integrated model. The fixed effects, percentage of variance explained by the model (marginal r-squared), percentage of variance explained by each index (semi-partial r-squared), and the AIC value of each model are also provided.

Table 4.16 Comparisons with the L2 Independent Model

Independent Models	***Marginal R^2***	***AIC***	***Indices***	***Semi-partial R^2***	***L2 Independent Model***
L2 Independent	13.0%	1146.7	L2 Word Naming RT	13.0%	
L1 Independent	11.0%	1162.1	ELP Word Naming RT CW	11.0%	r = 0.000, p = 0.50
Combined Independent	15.0%	1140.0	L2 Word Naming RT ELP Word Naming RT	4.0% 2.0%	r = 0.000, p = 0.50

Table 4.17 Comparisons with the L2 Integrated Model

Integrated Models	***Marginal R^2***	***AIC***	***Indices***	***Semi-partial R^2***	***L2 Integrated Model***
L2 Integrated	4.0%	1538.5	L2 Word Naming RT L2 Word Naming Accuracy	1.0% 2.0%	
L1 Integrated	6.0%	1530.9	ELP Word Naming RT CW ELP Word Naming Accuracy CW	6.0% 2.0%	r = 0.05, p = 0.09
Combined Integrated	7.0%	1551.5	ELP Word Naming RT CW L2 Word Naming Accuracy ELP Word Naming Accuracy CW	4.0% 1.0% 1.0%	r = –0.03, p = 0.18

Tables 4.16 and 4.17 summarize the answer to the second research question, which asked to what degree the L2 and L1 indices of word recognition explained writing quality. Both L2 and L1 models, either independent or integrated, were compatible (i.e., they explained a similar amount of variance in the scores) as suggested by the lack of statistical differences between models. Only word naming reaction time explained the independent scores. There was a positive relationship between independent scores and RT, suggesting that the test takers that used words that took longer to be processed either by L2 or L1 speakers of English earned higher scores. The integrated scores were explained both by reaction time and accuracy indices. Similar to the independent models, RT had a positive relationship with integrated scores. The L2 accuracy index had a negative relationship with test scores, whereas the L1 accuracy index had a positive relationship with scores. This means that when test takers used words that are more difficult to L2 users and easier to L1 users, their scores were higher in the integrated task (see discussion below for an alternative explanation). The answer to the second research question is that both the L2 and L1 indices were predictors of L2 writing, with reaction time predicting independent essay scores, and both reaction time and accuracy predicting integrated essay scores.

4.6 Discussion

The use of behavioral psycholinguistic data as benchmarks of lexical sophistication has been extensively used in the automatic assessment of L2 lexical proficiency (e.g., Berger, Crossley, & Kyle; 2019; Crossley & McNamara, 2009; Kyle et al., 2018). These indices have helped understand how word properties such as concreteness and associative context judged by human raters influence L2 production. Despite achievements in this area, the behavioral data used so far have come from L1 judgements made available in large databases, such as familiarity judgements (Coltheart, 1981) and word processing information from monolingual participants

(Balota et al., 2007). The reliance on L1 psycholinguistic data can be partially attributed to the lack of large L2 datasets. Study 3 sets out to address this gap in the lack of L2 psycholinguistic data by collecting word recognition information for 4,998 words, which was turned into L2 indices of word recognition. To validate these indices, they were compared to similar L2 and L1 word recognition databases and used as explanatory variables of human ratings of L2 writing quality.

In the first validation step, the convergent validity of the L2 word recognition indices was tested by comparing them with similar L1 word naming norms and L2 lexical decision norms. This step answered the first research question of Study 3, which asked how the L2 word recognition information compare to similar L1 and L2 word recognition norms. The comparison with the L1 word naming norms revealed a medium correlation for reaction time, which is the main index of lexical access (de Groot, 2011), and a low correlation for accuracy, which is related to the breadth of lexical knowledge. Studies on L1 and L2 word processing may help understand these correlations. L2 processing has been shown to be slower and less accurate than L1 processing (de Groot et al., 2002; Kaur, 2017). Scholars have attributed these differences to degree of exposure instead of differences in how bilingual and monolinguals process language (Brysbaert et al., 2000; Monaghan et al., 2017). Because exposure to input tends to be more limited among L2 users, with input being unnatural or modified (Assche et al., 2020), quantitative differences related to the depth and breadth of lexical knowledge are expected. These differences in linguistic exposure might explain the lack of a high correlation. In other words, the medium correlation with RT and low correlation with accuracy may be taken to support the notion that both the L1 and L2 word naming indices are measuring similar constructs

while reflecting quantitative differences in processing, which may be more prominent regarding the breadth of lexical knowledge.

Comparisons with the L2 lexical decision database from Berger, Crossley, and Skalicky (2019) showed similar trends: medium correlation for lexical decision RT z-scores and low correlations for accuracy scores. These findings can be related to differences between the word naming and lexical decision task and the subject pool included in this dissertation and Berger, Crossley, and Skalicky (2019). While both the word naming and lexical decision tasks tap on word recognition processes, important differences exist. In word naming, where production is necessary, phonological representations must be accessed, which requires extra processing time (de Groot, 2011). On the other hand, lexical decision tasks reflect greater semantic processing (de Groot et al., 2002). These differences in the types of representations that are mostly required by each task can explain why correlations were not higher. It is important to note that both tasks require lexical access, which entails the simultaneous activation of orthographic, phonological, and morpho-syntactic knowledge (Monaghan et al., 2017), but deeper activation of these levels for each task may occur. The second source of differences might be related to the participant population. The participants in Berger, Crossley, and Skalicky's (2019) study differed from the participants in this dissertation in important ways. In this dissertation, the participants were college students and primarily young adults with more limited, albeit diverse, exposure to English, whereas the participants in Berger, Crossley, and Skalicky (2019) had a more extensive range of linguistic experiences. Also, although Romance languages (e.g., Portuguese, Spanish, French) were the primary L1s spoken by the participants in both databases, proportionally, there was a greater representation of participants speaking Chinese and Korean as an L1 in the word naming database. These participant and task differences may explain the medium to low

correlation scores, while the overlap between the two tasks may be attributed to similarities in lexical access measurement.

The second validation step entailed the use of the L2 word recognition indices as explanatory variables of L2 writing quality. This step provided the answer to the second research question, which asked to what degree the L2 and L1 word recognition indices can explain human judgements of L2 writing. Essays from two TOEFL writing tasks (i.e., integrated and independent) were used. The results suggested that both the L2 and L1 indices explained part of the integrated and independent writing scores, with no statistical differences between them. In the independent model, the L2 indices explained 13% of the variance in scores, whereas the L1 indices explained 11% of the scores. In the integrated task, the L2 indices explained 4% of the scores, whereas the L1 indices explained 6% of the scores. More variance was explained when the indices were combined (i.e., 15% of the independent scores and 7% of the integrated scores).

In the independent models, the index reaction time, both from the L1 database (i.e., ELP) and the L2 word naming database, was the only significant predictor. Higher RT scores were associated with higher independent essay scores, meaning that proficient writers used more words that pose processing challenges to both L1 and L2 users. A low-scored and high-scored essay are featured in Appendix M, along with the individual output for the significant indices. The values that contributed to higher scores are highlighted in red. The lower-scored essay contained substantially more words with lower L2 RT scores as compared to the test-takers' average, including "keep," "stop," "smoke," and "bad." The high-scored essay included more words and a wider range of L2 RT scores, including "pursue," "engineering," "mechanical," and "complicated." The coverage of the L1 index, which includes RT information for approximately

40,000 words, was higher, but the pattern was the same: there were more sophisticated words in the high-scored essay.

In the integrated models, both RT and accuracy scores from both the L2 and L1 datasets helped explain essay scores. The effect of the RT scores in the integrated task was similar to its effect in the independent task; the writers who used words that posed processing challenges to both L1 and L2 users performed better. However, the effect of accuracy as measured by the L1 index was positive, whereas the effect of accuracy as measured by the L2 index was negative. In other words, writers that gave preference to words that were easier to L1 users but more difficult to L2 users received higher scores. A ceiling effect, which is particularly stronger in the L1 list, may have caused this differential effect. Overall, most test takers gave preference to words with high accuracy values as measured by the L2 accuracy index (mean = 0.973, Min = 0.946, Max = 0.991) and the L1 accuracy index (mean = 0.994, Min = 0.986, Max = 0.998). As illustrated in Appendix N, which features the individual output of a low-scored and high-scored integrated essay, the majority of the accuracy scores for the words in the essays was 1 (i.e., most words received a perfect accuracy score). Because half of the words in the ELP dataset received a perfect accuracy score (N = 20,088), longer essays, which tend to get higher scores, had an increased chance of receiving several scores of 1, giving more elaborated essays a misleading higher accuracy score. The L2 accuracy index, on the other hand, captured more processing differences among words than the L1 accuracy scores.[17] For example, the word "species," "wild," and "lecturer" had L2 accuracy scores of 0.76, 0.81, and 0.92 respectively (see individual output in Appendix N), but they had a perfect accuracy score in the ELP database. It seems, then,

[17] From the 4,998 words that overlapped between the two lists, 4061 of these words in the ELP list had a perfect accuracy score (accuracy = 1), whereas only 1970 of the words in the L2 list had a perfect accuracy score.

that the L2 accuracy indices may be better at differentiating word processing difficulties and may be more reliable word recognition indices.

A post-hoc analysis with the ELP subset that overlapped with the words available in the L2 database was performed for more direct comparisons with the L2 independent and integrated models. The new models are called L1 (ELP) overlapping models. The results are presented in Appendix O, which includes the correlations of the L1 overlapping indices with the TOEFL scores, model comparison statistics, and statistics of the independent and integrated model with the best fit. Different from the independent model with all ELP words, where RT was the only predictor, the overlapping ELP independent model retained two indices: accuracy and reaction time SD. These indices explained only 2% of the variance, 9% less than the complete model. Reaction time SD had a positive impact on the model, meaning that high scorers used words that present processing challenges for a subset of L1 users. Accuracy also had a positive effect on independent scores, similar to the ELP integrated model reported above. Appendix M, which also features the ELP overlapping indices, shows a similar high incidence of perfect scores in both the low-scored and the high-scored essay, corroborating the previous finding that this index may not be discriminating between more sophisticated and less sophisticated items. The ELP overlapping integrated model also retained the indices reaction time SD and accuracy and reported the same positive effect of accuracy on essay scores. A comparison with the ELP overlapping and L2 accuracy indices available in Appendix N confirms that the L2 accuracy index was distinguishing more words (i.e., there were more values below 1) and that most of the ELP scores for accuracy were 1. Therefore, the overlapping models provided further evidence that the L2 indices may be more valid representations of word difficulty among L2 writers.

The findings from the L2 writing models (i.e., the models that utilized the L2 word recognition indices) replicate previous research. Similar to Berger, Crossley, and Kyle (2019), who analyzed human ratings of lexical proficiency in speaking, high proficiency was related to the use of words that are more difficult to process, as measured by both RT and accuracy scores. Also, in line with Kyle et al. (2018), who investigated L2 writing, more proficient writers used words that are more difficult to process as measured by accuracy scores. However, contrary to Berger, Crossley, and Kyle (2019) and Kyle et al. (2018), who adopted the same L1 indices used here, the L1 accuracy models suggested that more proficient writers used words that are easier to process by L1 speakers. These differences may be related to the differences in mode (i.e., spoken versus written) in Berger, Crossley, and Kyle (2019) and the writing tasks (i.e., free write vs. high-stakes essay) in Kyle et al. (2018); however, a comparison of individual output is necessary to judge whether their studies did not suffer from the same ceiling effect in the accuracy scores.

Similar to Study 1 and 2, the variance explained by the lexical indices accounted for only a portion of the L2 writing scores. This finding may be related to the holistic nature of the rubric, as opposed to a rubric assessing lexical proficiency, and to the fact that other important factors related to writing proficiency were not accounted for, such as cohesion and completeness of response (Biber & Gray, 2013; Cumming et al., 2006). The lower variance that was explained in the integrated models may be related to the influence of the integrated words from the source into the texts, which affects the lexical scores, as demonstrated in Study 1 and Study 2.

4.7 Conclusion and Limitations

This study set out to validate indices of word recognition based on L2 word processing data. The results suggested that the L2 indices are valid representations of word processing difficulty that can be used in the automatic analysis of L2 texts. Particularly, the L2 indices have

been shown to be significant predictors of independent and integrated essay quality, as judged by human raters.

An argument can be made that L2 word recognition indices contain a larger range of processing information than the L1 word recognition indices, potentially being more representative of processing differences in L2 texts. As indicated by lexical processing studies, bilingual processing tends to present more variation (Brysbaert et al., 2017; Diependaele et al., 2013; Gollan et al., 2011), which is related to the differential experience that L2 users have with language (Johns et al., 2016). Monolingual processing, on the other hand, tends to reach processing thresholds, showing little differences for frequent items (Assche et al., 2020). This indicates that when data are collected from experienced language users, which is more common among L1 users, important developmental processing information may be lost. This issue was found with the L1 accuracy norms, which presented a higher ceiling effect than the L2 norms. It is possible that L2 processing data may be better suited to represent L2 word processing because it carries greater information related to different linguistic experiences.

A few limitations regarding word recognition indices should be made. While word naming tasks have been widely used to the study of word recognition, or lexical access, the information derived from this task is more directly related to comprehension instead of production, which requires the spontaneous retrieval of lexical items from memory (Gollan et al., 2011). Also, performance in word naming tasks can bypass recognition if respondents rely on the application of script-to-sound rules. It is worth noting, though, that this reliance is less relevant for English (de Groot, 2011). Finally, this study relied on words as a unit of analysis, which may be limited in its ability to represent lexical knowledge. Phrases such as n-grams and phrase frames have been claimed to be the basic unit of language (Sinclair, 2008); therefore, larger and

productive units may be better suited to gauge lexical sophistication. Notwithstanding these limitations, analyses of individual output and comparisons with L1 indices have suggested that the L2 word recognition norms are representative of the difficulty in processing among L2 writers of English.

5 CONCLUSION

A long-held tradition in psycholinguistics and applied linguistics has been to compare L2 production and processing behavior with those of L1 users (Klein, 1998; Ortega, 2016). While this tradition has helped us understand many L2 linguistic phenomena, some scholars have argued that it might have limited our assessment of multicompetence (Cook, 1991; Vaid & Meuter, 2017). For example, Vaid and Meuter (2017) criticized the selection of psycholinguistic stimuli based exclusively on L1 frequency lists, warning that this selection might not be relevant to the bilingual experience and affect the strength of the variables under investigation. Naismith et al. (2018) also questioned the use of L2 lexical proficiency measurements based on L1 corpora, arguing that L2 users, especially at beginning levels, have linguistic needs different from L1 users; therefore, assessments based exclusively on L1 benchmarks may have exacerbated differences or created gaps in studies. One of the reasons L2 researchers rely on L1 benchmarks is the lack of diverse and robust L2 automatic indices available to them. Indices based on L2 corpora may afford the opportunity to replicate analyses based on L1 indices to test the strength of past conclusions and new hypotheses regarding L2 production (Bestgen, 2017; Porte, 2012; Vaid & Meuter, 2017). This dissertation helped address the gap in the scarcity of L2 lexical benchmarks by developing and testing four types of automatic indices: lexical frequency, range, semantic context, and word recognition indices.

The validity of the L2 indices was tested by replicating previous applied linguistics and psycholinguistics analyses that have utilized L1 indices. Specifically, the new L2 indices were used to model L2 writing quality and L2 lexical processing, which helped test the validity of the L2 indices as benchmarks of lexical sophistication. These validation steps have been extensively used in studies testing new and improved lexical benchmarks (Adelman et al., 2006; Heuven et

al., 2014; Mandera et al., 2017). Two writing tasks were included to model L2 writing proficiency: independent and integrated writing tasks from the TOEFL iBT. The L2 indices were used to explain L2 writing proficiency as judged by expert human raters. Accuracy and reaction time scores from a lexical decision task performed by L2 users were used to model L2 lexical processing. The 3,338 words available in the L2 lexical decision dataset from Berger, Crossley, and Skalicky (2019) were analyzed using the L2 norms, which were used as explanatory variables of L2 lexical accuracy and reaction time scores. The same models were run using comparable L1 indices for comparison purposes. Combined models with both L1 and L2 indices were also run to test how the indices complement each other and to test whether the L2 indices surfaced as predictors when L1 indices were added as control variables. The results from each validation step are summarized below.

The independent models investigated the power of the L2 and L1 indices in predicting the scores of TOEFL independent essays, which consist of impromptu writing under time constraints. The time constraints and the lack of sources from which to extract ideas and vocabulary make independent essays good candidates to assess lexical knowledge. All L2 lexical sophistication indices tested in this dissertation surfaced as explanatory variables of L2 writing quality, and all suggested that proficient writers prefer more sophisticated words, lemmas, and lemma-based n-grams. The L2 semantic context indices were particularly predictive of the independent scores, explaining 18% of the variance. These findings suggest that proficient writers more frequently relied on lemmas that were less semantically rich (i.e., lemmas that appear in fewer semantic contexts) and lemmas that are more distinct (i.e., that occur in specific contexts). The L2 word recognition indices explained 13% of the independent scores and suggested that more proficient writers rely on words that are more difficult to process by L2

users of English. The L2 frequency and range indices explained 11% of the independent scores and suggested that proficient writers rely on bigrams that are more sophisticated. The L1 frequency and range indices explained 12% of the independent scores and suggested that more proficient writers gave preference to more sophisticated content lemmas. The L1 word recognition indices explained 11% of the scores in the complete model and 2% in the overlapping model (i.e., the models that only included word recognition information from ELP that overlapped with the L2 indices). The L1 RT and SD indices suggested that proficient L2 writers prefer words that pose processing challenges to L1 users. The accuracy scores suggested an opposite trend, but the effect of this index suffered from a ceiling effect; therefore, its results may be inaccurate. The L1 semantic context indices did not explain any variance in the independent scores. The L2 frequency and range indices were comparable to the L1 indices in terms of statistical power, but the L2 semantic context and L2 word recognition indices in the post-hoc analyses were more predictive of the independent scores (see discussion below for possible reasons).

The integrated models investigated the power of the L2 and L1 indices in predicting the scores of TOEFL integrated essays. Although this task is also constrained by time, test-takers are expected to rely on sources for ideas and vocabulary when summarizing and critiquing the sources. Therefore, the lexical items in the essays may not always reflect the writers' lexical proficiency because they are expected to borrow words from the sources. Not surprisingly, all indices, including the L1 indices, had a lower predictive power in this task compared to the independent task. The L2 semantic context indices explained 10% of the integrated scores, with proficient writers giving preference to lemmas that tend to be acquired later and are more semantically rich. The L2 word recognition indices explained 4% of the integrated scores, with

more proficient writers using words that are more difficult and produced less accurately by L2 users of English. The L2 frequency indices also explained 4% of the integrated scores, with more sophisticated trigrams being associated with higher scores. The L1 indices showed similar trends. The L1 frequency and range indices explained 1% of the integrated scores and they suggested that more proficient writers gave preference to more sophisticated n-grams. The L1 word recognition indices explained 6% of the integrated scores in the complete models and 4.5% in the overlapping models. They suggested that proficient writers preferred words that pose processing challenge as indexed by RT and SD scores. The L1 semantic context indices explained 5% of the integrated scores and they suggested that more proficient writers gave preference to less distinct and more semantically rich lemmas. Both the L2 and L1 frequency and word recognition indices had a statistically similar explanatory power; however, the L2 semantic context indices were statistically stronger. Additionally, analyses of individual input suggested that the L1 word naming accuracy index suffered from a ceiling effect and that its power may be related to text length instead of processing information.

The L2 lexical processing models included the L2 and L1 frequency, range, and semantic context norms reported in Study 1 and 2. The L2 range index explained 13.5% of the reaction time scores and 14% of the accuracy scores. More sophisticated words (i.e., words that occurred in fewer contexts) were more difficult to process. The L2 semantic context indices explained 9% of the reaction time scores and 8.6% of the accuracy scores. The L1 range indices explained 15% of the variance in RT scores and 11% of the variance in accuracy scores, and the L1 semantic context indices explained less than 1% of the variance in processing scores. Overall, the models suggested that lemmas that are more semantically rich, are less distinct, and have fewer distinct relationships are processed faster. The L2 semantic context indices were stronger predictors of

L2 processing than the L1 indices, which explained less than 1% of the reaction time and accuracy scores; however, the L1 frequency and range were compatible with the L2 indices in explanatory power.

Overall, the results suggest that both the L1 and L2 frequency and context diversity indices are successful predictors of L2 writing and lexical processing, while the L2 word recognition indices may contain richer and more varied processing information that may better inform L2 lexical proficiency. The strength of the L2 semantic context indices found in this dissertation needs to be tested against other L1 indices, which, to the best of the author's knowledge, are not freely available and perhaps have not yet been developed. Any statistical differences between the L1 and L2 semantic benchmarks reported in this dissertation may be due to the limitations of the TASA corpus used in developing the L1 benchmarks. Notwithstanding the limitations in the comparisons, the L2 semantic indices were powerful predictors, especially in explaining independent essay scores. Additional possible reasons for the differences in the L2 and L1 indices are provided below. Due to the potential confounding effect coming from the integrated words in the integrated essay task, a focus is given to the independent models and L2 lexical processing models in explaining differential effects between the L2 and L1 indices.

Frequency and context diversity indices are taken as a proxy of the linguistic experience that language users have (Balota & Chumbley, 1985; Ellis, 2002). No single corpora can account for the totality of experiences someone has with language; therefore, an array of indices from different domains (e.g., academic, fiction) and modes (i.e., written and spoken) have been developed (Brysbaert & New, 2009; Kyle et al., 2018). Thus, it is not surprising that the L2 frequency and range indices were not stronger than the L1 indices as the L2 indices represent the indirect experience of some L2 users learning through writing; that is, they represent the

language from classroom tasks produced by language learners. The results should not be taken as evidence that EF-CAMDAT and COCA Fiction represent the same linguistic experience, though. As discussed in Study 1, almost half of the words in the top 1,000 EF-CAMDAT were not in the top 1,000 COCA Fiction, and correlations between the EF-CAMDAT and COCA Fiction ranged from .64 for single lemmas to .38 for trigrams. These differences caused an interesting trade-off effect of the indices on the L2 writing models. The models suggested that the n-grams indices from EF-CAMDAT were stronger predictors of essay scores, whereas single-lemma indices from COCA Fiction were more predictive. Analysis of individual output indicated that EF-CAMDAT had a higher coverage for bigrams, capturing more sophisticated word combinations than COCA Fiction. In the example in Appendix C, n-grams such as "be emphasize(d)," "personal development," "future profession," "external job," "future employer," "financial independence," "high demand," and "secure income," all with a lower range score as indexed by EF-CAMDAT had no counterpart in COCA Fiction. This might be due to the domain of each corpora. In COCA Fiction, topics related to career choice are much less likely to occur than in EF-CAMDAT, which includes a variety of academic and job-related tasks common in many English classrooms. Another possible explanation for the higher coverage of EF-CAMDAT indices is the lower cut-off point adopted in this dissertation (i.e., 2), which might have given an advantage for this corpus regarding n-gram coverage, increasing its explanatory power. It is worth noting, however, that models with a higher cut-off point set at 5, which are not reported here due to space constraints, showed a similar trend: n-grams from EF-CAMDAT were stronger predictors than n-grams from COCA Fiction. Regarding the advantage of content lemmas from COCA Fiction, analysis of individual output also revealed a domain effect. Some lemmas that are common in classroom tasks were indexed as less sophisticated by EF-CAMDAT, including "career,"

"opinion," "education," "market," "achieve," "financial," and "goal," which were indexed as more sophisticated by COCA Fiction. It is possible that essay raters consider linguistic knowledge beyond the domain of the classroom, thus judging lemmas like those as more sophisticated. This domain effect explains previous seemingly contradictory findings regarding the effect of n-grams in essay scores. Monteiro et al. (2020), for example, found that proficient writers gave preference to less sophisticated n-grams as indexed by COCA Academic in independent TOEFL essays. Their findings suggest that using common n-grams found in professional academic writing is indicative of high proficiency, whereas the results from this dissertation suggest that, when other domains are considered, more sophisticated n-grams are preferred.

Another interesting trade-off regarding the effect of the frequency and range norms was observed in the L2 lexical processing models. While COCA Fiction explained more of the reaction time scores (15% versus 13.5%), EF-CAMDAT explained more of the accuracy scores (14% versus 11.4%). Corpus domain may also explain this finding. Appendix F shows that, overall, indices with lower range as indexed by both L1 and L2 indices were processed slower and less accurately; however, a few differences were found. The words that did not follow this pattern included the efficiently-processed words "imagination," "cow," "snake," "pink," "heaven," "doll," and "spider," which were indexed as more sophisticated by EF-CAMDAT and as less sophisticated by COCA Fiction. These are words that most L1 users learn at a relatively early age, being featured more frequently in works of fiction, but they may not be common in average adult L2 classrooms. The L2 users who participated in the lexical decision task, most of whom lived in an English-speaking country, were probably frequently exposed to these words. Interestingly, similar words were processed with less accuracy by some participants, including

"hip," "fist," "iron," "dust," "waist," "patch," "faint," and "bell," indexed by EF-CAMDAT as more sophisticated, but as less sophisticated by COCA Fiction. It is possible that some participants in Berger, Crossley, and Skalicky (2019) had a domain-specific knowledge of English and did not know some words that are common in works of fiction. Because accuracy scores are more sensitive to inaccurate responses than reaction time scores are to slower responses, differences in experiences among L2 users may have had a larger impact on accuracy scores. A follow-up study investigating the words that were indexed differently by COCA Fiction and EF-CAMDAT as outcome variables, using age of arrival and other benchmarks related to linguistic experiences as explanatory variables, could help clarify this trade-off effect.

Semantic context indices are also taken to represent the experience with input (Landauer et al., 1998; McDonald & Shillcock, 2001); however, different from frequency and range indices, they consider distributional properties based on word co-occurrence. The strength of the EF-CAMDAT indices over the TASA indices found in this dissertation may be due to the limitations in TASA and to the linguistic experience that each corpus represents. The TASA indices contained context information for 4,487 lemmas, less than one third than EF-CAMDAT (i.e., 16,031 lemmas). This alone may have reduced its power to analyze relevant information from texts; however, even in the combined L2 processing models, where only the overlapping lemmas (i.e., the lemmas that were both in TASA and EF-CAMDAT) were considered, the TASA indices performed poorly, explaining approximately 1% of the RT and accuracy scores. Correlations between the TASA and EF-CAMDAT indices were low (see Appendix G), suggesting that the indices do not represent the same semantic knowledge. In fact, analyses of individual output suggested that, for many lemmas, the index scores from EF-CAMDAT and TASA were in opposite directions. The individual output for the L2 word processing models

(Appendix K), which features the index average of all cosines from both corpora illustrates these differences. The lemmas indexed as more semantically rich by EF-CAMDAT were more clearly concentrated among the words processed faster (see table with RT results), whereas the lemmas indexed by TASA as semantically rich were unevenly spread. This finding might be due to the differences in corpus representativeness. As argued in Chapter 3, Study 2, TASA has been extensively used in studies to represent the American student experience with printed materials in school (Dascălu et al., 2014, 2016; Johns & Jones, 2008; Landauer & Dumais, 1997). Additionally, the texts in TASA are edited, instead of naturally produced by students at different school levels. EF-CAMDAT, on the other hand, represents the learning experience of many foreign language learners across the world and is much closer to natural production than TASA. Computational differences may have also caused the differences between the indices. Unfortunately, due to the lack of information regarding how the TASA indices available in TAALES were developed (see Kyle et al., 2018), it was not possible to compare computational differences.

Word recognition indices are taken to represent word processing difficulties, which can serve as a benchmark of lexical sophistication (Berger et al. 2019; Kyle et al., 2018). Even though the L2 word recognition indices were not more predictive than the L1 indices when the entire ELP dataset was considered, there was a clear advantage for the L2 indices when only the overlapping words were considered in the independent models (i.e., the L2 indices explained 13% of the independent scores versus 2% by the L1 indices). These differences may be due to the higher variability in the L2 RT and accuracy scores, which seem to carry greater information related to different linguistic experiences. As presented in Study 3, the range of RT and accuracy scores for the L2 indices was much higher than the range for the L1 indices. Additionally, for the

4,998 overlapping words, 4,061 of these had a perfect accuracy score (accuracy = 1) in the ELP list, whereas only 1,970 of the words in the L2 list had a perfect accuracy score. This higher ceiling effect in the L1 indices resulted in an inaccurate positive relationship between the L1 accuracy index and integrated scores. Based on these findings, it seems that L2 processing data may be a better source of word difficulty for L2 users for capturing processing difficulties that may be reduced due to accumulated experience found for many L1 users. Larger L2 datasets from mega-studies may be needed to test if this effect persists with more word information.

This dissertation has also brought into light some issues regarding the use of L2 corpora and L2 behavioral data. As discussed in previous chapters, a definition of a threshold for word inclusion in frequency and range indices needs to be carefully considered and tested. This may be especially true when learner corpora, which may contain more misspellings, are considered. While, for the present study, quantitative analyses revealed that a less conservative threshold was appropriate and afforded the analyses of more sophisticated n-grams, a more conservative threshold might be more appropriate for other purposes such as material development or high-stakes assessment. The inclusion of non-standard forms such as "belive," which is highly frequent in EF-CAMDAT, also needs to be considered. Another issue that is important to consider is whether to include texts written by less proficient L2 users. From an analytical standpoint, lower-level texts may be difficult to process and parse (Meurers & Dickinson, 2017), but they may be an invaluable source of information for material development (Naismith et al., 2018) and for analyses of L2 development (Crossley, Skalicky, et al., 2019). A solution might be to develop benchmarks by language levels as in Monteiro et al. (2020) and tools that allow the researcher to manipulate thresholds and the incorporation or exclusion of non-standard forms. The inclusion of different proficiency levels in behavioral studies is another point to consider. L2

users at earlier stages of language development have lexical representations that are still closely attached to their L1(s) (Assche et al., 2020), bringing noise (e.g., outliers) to datasets. There are also practical issues related to the time and length of behavioral tasks, in that including participants of multiple language levels would require tasks to be adapted to different L2 users. Again, a solution might be to develop processing benchmarks by language level.

In addition to the limitations highlighted in each chapter, some general limitations need to be considered. The most important limitation is that this dissertation treated L2 users as an undifferentiated category. The same way that an L1 corpus or dataset cannot encompass all linguistic and behavioral experiences that L1 users have, the indices developed for this dissertation are limited by the corpus and the participants selected for the word naming task. Specifically, EF-CAMDAT represents the experience of formal language learning, and the word naming norms reflect the experience of L2 users who have lived part of their lives in an English-speaking country. The limitations of L2 indices may be even more relevant than the limitations of L1 indices, as variability is inherent to the bilingual experience (Vaid & Meuter, 2017), requiring multiple representations. Another important limitation is that this study engaged in some of the same deficit practices as other studies by comparing the L2 benchmarks to L1 benchmarks, and by selecting words for the word naming task from an L1 dataset (i.e., ELP) and corpus (i.e., COCA Spoken). Possibly, a fairer comparison would have been to test the indices against other L2 norms, which would have allowed an investigation of how L2 indices differ from one another. The experimental material for the word naming task could also have benefited from having used L2 frequency indices as a benchmark. Another limitation of EF-CAMDAT is that some of the lexical items may have been copied from the *Englishtown* tasks; that is, some of the language in EF-CAMDAT may not be naturally produced. Lastly, the validation steps

included in this dissertation were restricted to cross-sectional analyses and were not tested in the presence of confounding variables. Future research may be necessary to test the potential of the L2 indices for explaining longitudinal L2 data, especially in comparison to other L2 indices, to further test their strength to model L2 behavior. Also, the L2 indices tested separately in the three studies need to be tested in combination and in models that control for confounding lexical sophistication benchmarks such as familiarity, concreteness, and age of acquisition. This would allow for an investigation of the predictive power of the indices beyond what other variables can explain.

Notwithstanding the limitations, this dissertation has suggested that L2 indices that measure frequency, context diversity, semantic context, and word recognition can provide additional information about L2 linguistic experiences and serve as lexical sophistication benchmarks for multiple text analyses. More importantly, it has shown that L2 indices can bring lexical sophistication and processing information that may be unique to this type of index. It is worth pointing out that the L1 indices used here, especially the frequency and range indices, were also successful predictors of L2 writing and processing, and they should continue to be used in L2 text analysis to represent the experience with input that L2 users have with L1 input. In other words, the importance of the L2 indices does not invalidate the importance of L1 indices, but only emphasize the need for multiple and diverse lexical sophistication indices.

This dissertation has also been a step towards representing lexical sophistication beyond the traditional frequency norms, by adding semantic context and word recognition indices, and it opens possibilities for future research on many levels. Firstly, the replication of studies that have used L1 norms may be necessary, especially those that have relied on L1 indices for the selection of experimental material (Vaid & Meuter, 2017). Research on L2 lexical proficiency should also

be reassessed based on L2 benchmarks. Finally, more L2 indices need to be developed. We still have a long way to develop benchmarks that represent the different types of L2 linguistic experiences, including, for example, norms that represent English as a Lingua Franca. Specifically, other L2 indices can be developed from ELF corpora such as ELFA (Mauranen et al., 2010) and EAP corpora such as ICLE (Granger et al., 2009). Also, large L2 spoken corpora should be collected to develop indices that can provide a naturalist sample of L2 production to be used to analyze L2 speaking. Ideally, a range of L2 indices should be developed from different domains (e.g., academic writing, academic speaking, everyday conversations) similar to the different indices that are available from L1 data.

The practical applications are also noteworthy. Similar to how the L1 indices have been utilized, the L2 indices can be used to generate models that can serve to assess writing (McNamara et al., 2015; Ramineni & Williamson, 2013) and readability (Dascălu et al., 2012, 2013), to refine frequency lists that serve as benchmarks for L2 teaching (Coxhead, 2000; West, 1953), and to adapt classroom materials (Kim & Monteiro, 2019; Monteiro & Kim, in press). In general, the indices can be used to automatically analyze discourse to explore human behavior, psychological processes, and cognitive processes (McNamara et al., 2017), increasing the opportunities that NLP can offer for robust and diverse natural language investigations (Meurers, 2013).

Funding

This dissertation has been partially funded by the ETS "Small Grants for Doctoral Research in Second or Foreign Language Assessment" and the ALRC "Research Support Grant."

REFERENCES

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*(2), 81–92. https://doi.org/10.1016/j.esp.2011.08.004

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*(9), 814–823. https://doi.org/10.1111/j.1467–9280.2006.01787.x

Altszyler, E., Sigman, M., & Slezak, D. F. (2018). Corpus specificity in LSA and Word2vec: The role of out-of-domain documents. *Proceedings of the 3rd Workshop on Representation Learning for NLP*, 1–10. http://arxiv.org/abs/1712.10054

Arumugam, R., & Shanmugamani, R. (2018). *Hands-On Natural Language Processing with Python: A practical guide to applying deep learning architectures to your NLP applications*. Packt Publishing Ltd.

Assche, E. van, Brysbaert, M., & Duyck, W. (2020). Bilingual lexical access. In R. R. Heredia & A. B. Cielicka (Eds.), *Bilingual lexical ambiguity resolution* (pp. 42–67). Cambridge University Press.

Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, *5*(3), 436–461. https://doi.org/10.1075/ml.5.3.10baa

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Balota, D. A., & Chumbley, J. I. (1985). The locus of word-frequency effects in the pronunciation task: Lexical access and/or production? *Journal of Memory and Language*, *24*(1), 89–106. https://doi.org/10.1016/0749–596X(85)90017–8

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual Word Recognition of Single-Syllable Words. *Journal of Experimental Psychology: General*, *133*(2), 283–316. https://doi.org/10.1037/0096–3445.133.2.283

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), *Visual word recognition: Models and methods, orthography and phonology* (pp. 90–114). Psychology Press.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Baron, A., Rayson, P., & Archer, D. (2009). Word frequency and key word statistics in corpus linguistics. *Anglistik*, *20*(1), 41–67.

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

Bates, E., Burani, C., D'Amico, S., & Barca, L. (2001). Word reading and picture naming in Italian. *Memory & Cognition*, *29*(7), 986–999. https://doi.org/10.3758/BF03195761

Berger, C. M., Crossley, S. A., & Kyle, K. (2017). Using novel word context measures to predict human ratings of lexical proficiency. *Journal of Educational Technology & Society*, *20*(2), 201–212.

Berger, C. M., Crossley, S. A., & Kyle, K. (2019). Using native-speaker psycholinguistic norms to predict lexical proficiency and development in second-language production. *Applied Linguistics*, *40*(1), 22–42. https://doi.org/10.1093/applin/amx005

Berger, C. M., Crossley, S. A., & Skalicky, S. (2019). Using lexical features to investigate second language lexical decision performance. *Studies in Second Language Acquisition*, *41*, 911–935. https://doi.org/doi:10.1017/S0272263119000019

Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System*, *69*, 65–78. https://doi.org/10.1016/j.system.2017.08.004

Bestgen, Y., Lories, G., & Thewissen, J. (2010). Using latent semantic analysis to measure coherence in essays by foreign language learners? *In Proceedings of 10th International Conferences Journée d'Analyse Statistique Des Données Texuelle (JADT2010)/Bolasco*, 1–12.

Bialystok, E. (2009). Bilingualism: The good, the bad, and the indifferent. *Bilingualism: Language and Cognition*, *12*(1), 3–11. https://doi.org/10.1017/S1366728908003477

Biber, D., Conrad, S., & Leech, G. (2002). *Student grammar of spoken and written English*. Pearson Education Limited.

Biber, D., & Gray, B. (2013). Discourse characteristics of writing and speaking task types on the TOEFL Ibt® test: A lexico-grammatical analysis. *ETS Research Report Series*, *2013*(1), i–128. https://doi.org/10.1002/j.2333–8504.2013.tb02311.x

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Botarleanu, R.-M. (2020, February 12). [Personal communication].

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-Academic-Purposes speaking tasks* (Vol. 2005). ETS. https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333–8504.2005.tb01982.x

Brysbaert, M., Lagrou, E., & Stevens, M. (2017). Visual word recognition in a second language: A test of the lexical entrenchment hypothesis with lexical decision times. *Bilingualism: Language and Cognition*, *20*(3), 530–548. https://doi.org/10.1017/S1366728916000353

Brysbaert, M., Lange, M., & Wijnendaele, I. V. (2000). The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition: Further evidence from the Dutch language. *European Journal of Cognitive Psychology*, *12*(1), 65–85. https://doi.org/10.1080/095414400382208

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior research methods*, *44*(4), 991-997.

Bychkovska, T., & Lee, J. J. (2017). At the same time: Lexical bundles in L1 and L2 university student argumentative writing. *Journal of English for Academic Purposes*, *30*, 38–52. https://doi.org/10.1016/j.jeap.2017.10.008

Chater, N., & Christiansen, M. H. (1999). Connectionism and natural language processing. In S. C. Garrod & M. J. Pickering (Eds.), *Language processing* (pp. 113–132). Psychology Press.

Colombo, L., Pasini, M., & Balota, D. A. (2006). Dissociating the influence of familiarity and meaningfulness from word frequency in naming and lexical decision performance. *Memory & Cognition*, *34*(6), 1312–1324. https://doi.org/10.3758/BF03193274

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, *1*, 497–505.

Cook, V. J. (1991). The poverty-of-the-stimulus argument and multicompetence. *Interlanguage Studies Bulletin (Utrecht)*, *7*(2), 103–117. https://doi.org/10.1177/026765839100700203

Cook, V. J. (1992). Evidence for Multicompetence. *Language Learning*, *42*(4), 557–591.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, *23*(4), 397–423. https://doi.org/10.1016/j.esp.2003.12.001

Cortese, M. J., & Schock, J. (2013). Imageability and age of acquisition effects in disyllabic word recognition. *The Quarterly Journal of Experimental Psychology*, *66*(5), 946–972. https://doi.org/10.1080/17470218.2012.722660

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Crossley, S. A., Kyle, K., Allen, L. K., Guo, L., & McNamara, D. S. (2014). Linguistic microfeatures to predict L2 writing proficiency: A case study in automated writing evaluation. *The Journal of Writing Assessment*, *7*(1). https://eric.ed.gov/?id=ED585968

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, *51*(1). https://doi.org/10.3758/s13428–018–1142–4

Crossley, S. A., & McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, *18*(2), 119–135. https://doi.org/10.1016/j.jslw.2009.02.002

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*(2), 115–135.

Crossley, S. A., & Salsbury, T. (2010). Using lexical indices to predict produced and not produced words in second language learners. *The Mental Lexicon*, *5*(1), 115-147.

Crossley, S. A., Salsbury, T., McNamara, D., & Jarvis, S. (2010). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561–580.

Crossley, S. A., Salsbury, T., & Mcnamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, *36*(5), 570–590. https://doi.org/10.1093/applin/amt056

Crossley, S. A., Salsbury, T., Titak, A., & McNamara, D. (2014). Frequency effects and second language lexical acquisition: Word types, word tokens, and word production. *International Journal of Corpus Linguistics*, *19*(3), 301–332. https://doi.org/10.1075/ijcl.19.3.01cro

Crossley, S. A., Skalicky, S., Kyle, K., & Monteiro, K. (2019). Absolute frequency effects in second language lexical acquisition. *Studies in Second Language Acquisition*, *41*(4), 721–744. https://doi.org/10.1017/S0272263118000268

Crossley, S. A., Subtirelu, N., & Salsbury, T. (2013). Frequency effects or context effects in second language word learning: What predicts early lexical production? *Studies in Second Language Acquisition*, *35*(4), 727–755.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.

Cuetos, F., & Barbón, A. (2006). Word naming in Spanish. *European Journal of Cognitive Psychology*, *18*(3), 415–436. https://doi.org/10.1080/13594320500165896

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., & Jamse, M. (2006). Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL®. *ETS Research Report Series – Wiley Online Library*. https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333–8504.2005.tb01990.x

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, *10*(1), 5–43. https://doi.org/10.1016/j.asw.2005.02.001

Dabbagh, A., & Enayat, M. J. (2019). The role of vocabulary breadth and depth in predicting second language descriptive writing performance. *The Language Learning Journal*, *47*(5), 575–590. https://doi.org/10.1080/09571736.2017.1335765

Dascălu, M., Dessus, P., Bianco, M., Trăușan-Matu, S., & Nardy, A. (2014). Mining Texts, Learner Productions and Strategies with ReaderBench. In A. Peña-Ayala (Ed.), *Educational Data Mining: Applications and Trends* (pp. 345–377). Springer International Publishing. https://doi.org/10.1007/978–3–319–02738–8_13

Dascălu, M., Dessus, P., Trăușan-Matu, Ş., Bianco, M., & Nardy, A. (2013). ReaderBench, an environment for analyzing text complexity and reading strategies. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial Intelligence in Education* (pp. 379–388). Springer. https://doi.org/10.1007/978–3–642–39112–5_39

Dascălu, M., McNamara, D. S., Crossley, S., & Trausan-Matu, S. (2016, March 5). Age of exposure: A model of word learning. *Thirtieth AAAI Conference on Artificial Intelligence*. https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11960

Dascălu, M., Trausan-Matu, S., & Dessus, P. (2012). Towards an integrated approach for evaluating textual complexity for learning purposes. In E. Popescu, Q. Li, R. Klamma, H. Leung, & M. Specht (Eds.), *Advances in Web-Based Learning—ICWL 2012* (pp. 268–278). Springer. https://doi.org/10.1007/978–3–642–33642–3_29

Davies, M. (2008). *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*. Available online at https://corpus.byu.edu/coca/.

de Groot, A. M. B. (2011). Comprehension processes: Word recognition and sentence processing. In *Language and cognition in bilinguals and multilinguals: An introduction* (pp. 155–220). Taylor & Francis Group. http://ebookcentral.proquest.com/lib/gsu/detail.action?docID=614735

de Groot, A. M. B., Borgwaldt, S., Bos, M., & van den Eijnden, E. (2002). Lexical decision and word naming in bilinguals: Language effects and task effects. *Journal of Memory and Language*, *47*(1), 91–124. https://doi.org/10.1006/jmla.2001.2840

Dell, G. S., Chang, F., & Griffin, Z. M. (1999). Connectionist models of language production: Lexical access and grammatical encoding. *Cognitive Science*, *23*(4), 517–542. https://doi.org/10.1207/s15516709cog2304_6

Diependaele, K., Lemhöfer, K., & Brysbaert, M. (2013). The word frequency effect in first- and second-language word recognition: A lexical entrenchment account. *The Quarterly Journal of Experimental Psychology*, *66*(5), 843–863.

Dijkstra, T., & Heuven, W. J. B. van. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, *5*(3), 175–197. https://doi.org/10.1017/S1366728902003012

Dijkstra, T., Jaarsveld, H. V., & Brinke, S. T. (1998). Interlingual homograph recognition: Effects of task demands and language intermixing. *Bilingualism: Language and Cognition*, *1*(1), 51–66. https://doi.org/10.1017/S1366728998000121

Dijkstra, T., Wahl, A., Buytenhuijs, F., Halem, N. V., Al-Jibouri, Z., Korte, M. D., & Rekké, S. (2019). Multilink: A computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition*, *22*(4), 657–679. https://doi.org/10.1017/S1366728918000287

Dikli, S. (2006). Automated essay scoring. *Turkish Online Journal of Distance Education-TOJDE, 7*(1), 49–62.

Duyck, W., Vanderelst, D., Desmet, T., & Hartsuiker, R. J. (2008). The frequency effect in second-language visual word recognition. *Psychonomic Bulletin & Review*, *15*(4), 850–855. https://doi.org/10.3758/PBR.15.4.850

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188. https://doi.org/10.1017/S0272263102002024

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). *Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of construction grammar*. Wiley.

Enright, M. K., & Tyson, E. (2008). Validity evidence supporting the interpretation and use of TOEFL iBT scores. *TOEFL IBT Research Insight, 4*, 1–21.

Fellbaum, C. (1998). Towards a representation of idioms in WordNet. *Usage of WordNet in Natural Language Processing Systems*, 52–57. https://www.aclweb.org/anthology/W98–0707

Fender, M. (2003). English word recognition and word integration skills of native Arabic- and Japanese-speaking learners of English as a second language. *Applied Psycholinguistics*, *24*(2), 289–315. https://doi.org/10.1017/S014271640300016X

Ferrand, L., Brysbaert, M., Keuleers, E., New, B., Bonin, P., Méot, A., Augustinova, M., & Pallier, C. (2011). Comparing word processing times in naming, lexical decision, and progressive demasking: Evidence from Chronolex. *Frontiers in Psychology*, *2*, 1-10. https://doi.org/10.3389/fpsyg.2011.00306

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. In *Studies in Linguistic Analysis* (pp. 1–32). Philological Society. https://ci.nii.ac.jp/naid/10020680394/

Foltz, P. W. (2007). Discourse coherence in LSA. In T. K. Landauer, D. S. McNamara, D. Simon, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 167–184). Erlbaum.

Friginal, E., Li, M., & Weigle, S. C. (2014). Revisiting multiple profiles of learner compositions: A comparison of highly rated NS and NNS essays. *Journal of Second Language Writing*, *23*, 1–16. https://doi.org/10.1016/j.jslw.2013.10.001

Garner, J., Crossley, S., & Kyle, K. (2019). N-gram measures and L2 writing proficiency. *System*, *80*, 176–187. https://doi.org/10.1016/j.system.2018.12.001

Gass, S. M., & Mackey, A. (2002). Frequency effects and second language acquisition: A complex picture? *Studies in Second Language Acquisition*, *24*(2), 249–260. https://doi.org/10.1017/S0272263102002097

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition, 7*(2), 95–112. https://doi.org/10.3758/BF03197590

Gollan, T. H., & Acenas, L.-A. R. (2004). What is a TOT? Cognate and translation effects on tip-of-the-tongue states in Spanish-English and Tagalog-English bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 246–269. https://doi.org/10.1037/0278–7393.30.1.246

González, M. C. (2017). The contribution of lexical diversity to college-level writing. *TESOL Journal*, *8*(4), 899–919. https://doi.org/10.1002/tesj.342

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. https://doi.org/10.3758/BF03195564

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). *International corpus of learner English*. Presses universitaires de Louvain

Green, C. (2012). A computational investigation of cohesion and lexical network density in L2 writing. *English Language Teaching*, *5*(8), 57–69.

Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software, 17*(1), 1-27.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*(3), 218–238. https://doi.org/10.1016/j.asw.2013.05.002

Hamrick, P., & Pandža, N. B. (2020). Contributions of semantic and contextual diversity to the word frequency effect in L2 lexical access. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *74*(1), 25–34. https://doi.org/10.1037/cep0000189

Heatley, A., Nation, P., & Coxhead, A. (2002). *Range and frequency programs*. http://www.victoria.ac.nz/lals/staff/paul-nation.aspx .

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, *21*(2), 303–317. https://doi.org/10.1017/S0272263199002089

Heuven, W. J. B. van, Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Hoey, M. (2005). *Lexical Priming: A new theory of language*. Routledge.

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718–730. https://doi.org/10.3758/s13428–012–0278-x

Huang, Y., Geertzen, J., Baker, R., Korhonen, A., & Alexoupoulou, T. (2017). *The EF Cambridge Open Language Database (EFCAMDAT)*. https://corpus.mml.cam.ac.uk/efcamdat2/public_html/EFCamDat-Intro_release2.pdf

Jaeger, B. (2016). Package 'r2glmm.' *R Package*, 1–9.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Linear model selection and regularization. In *An Introduction to Statistical Learning* (Vol. 103, pp. 203–264). Springer. https://doi.org/10.1007/978–1–4614–7138–7

Jamieson, R. K., Avery, J. E., Johns, B. T., & Jones, M. N. (2018). An instance theory of semantic memory. *Computational Brain & Behavior*, *1*(2), 119–136. https://doi.org/10.1007/s42113–018–0008–2

Jared, D., & Kroll, J. F. (2001). Do bilinguals activate phonological representations in one or both of their languages when naming words? *Journal of Memory and Language*, *44*(1), 2–31. https://doi.org/10.1006/jmla.2000.2747

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*(1), 57–84. https://doi.org/10.1191/0265532202lt220oa

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing*, *12*(4), 377–403. https://doi.org/10.1016/j.jslw.2003.09.001

Johns, B. T., Gruenenfelder, T. M., Pisoni, D. B., & Jones, M. N. (2012). Effects of word frequency, contextual diversity, and semantic distinctiveness on spoken word recognition. *The Journal of the Acoustical Society of America*, *132*(2), 74–80.

Johns, B. T., & Jones, M. N. (2008). Predicting word-naming and lexical decision times from a semantic space model. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *30*.

Johns, B. T., Sheppard, C. L., Jones, M. N., & Taler, V. (2016). The role of semantic diversity in word recognition across aging and bilingualism. *Frontiers in Psychology*, *7*, 1–11. https://doi.org/10.3389/fpsyg.2016.00703

Johnson, M. D., Acevedo, A., & Mercado, L. (2016). Vocabulary knowledge and vocabulary use in second language writing. *TESOL Journal*, *7*(3), 700–715. https://doi.org/10.1002/tesj.238

Jones, M. N., Dye, M., & Johns, B. T. (2017). Context as an Organizing Principle of the Lexicon. *Psychology of Learning and Motivation*, *67*, 239–283. https://doi.org/10.1016/bs.plm.2017.03.008

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *66*(2), 115–124. https://doi.org/10.1037/a0026727

Kameda, N. (1992). "Englishes" in cross-cultural business communication. *The Bulletin of the Association for Business Communication*, *55*(1), 3–8. https://doi.org/10.1177/108056999205500102

Kaur, S. (2017). Word naming in Bodo–Assamese bilinguals: The role of semantic Context, cognate status, second language age of acquisition and proficiency. *Journal of Psycholinguistic Research*, *46*(5), 1167–1186. https://doi.org/10.1007/s10936–017–9488–9

Kerkhofs, R., Dijkstra, T., Chwilla, D. J., & de Bruijn, E. R. A. (2006). Testing a model for bilingual semantic priming with interlingual homographs: RT and N400 effects. *Brain Research*, *1068*(1), 170–183. https://doi.org/10.1016/j.brainres.2005.10.087

Kim, Y., & Monteiro, K. R. (2019). The effect of input characteristics on students' perception of task difficulty and their comprehension of authentic listening tasks. In S. Sato & S. Loewen (Eds.), *Evidence-based second language pedagogy: A collection of instructed second language acquisition studies* (pp. 240–260). Routledge.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An Associative thesaurus of English and its computer analysis. In A. J. Aitken, R. W. Bailey, & N. Hamilton-Smith (Eds.), *The computer and literary studies* (pp. 153–165). Edinburgh University Press.

Klein, W. (1998). The contribution of Second Language Acquisition research. *Language Learning*, *48*(4), 527–549. https://doi.org/10.1111/0023–8333.00057

Kristofferson, A. B. (1957). Word recognition, meaningfulness, and familiarity. *Perceptual and Motor Skills*, *7*, 219–220. https://doi.org/10.2466/PMS.7..219–220

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428–012–0210–4

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). LmerTest: Tests in linear mixed effects models. R package version 2.0–20. *Journal of Statistical Software*, *82*(13), 1–30. https://doi.org/10.18637/jss.v082.i13

Kyle, K. (2020). The relationship between features of source text use and integrated writing quality. *Assessing Writing*, *45*, 100467.

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, *34*, 12–24. https://doi.org/10.1016/j.jslw.2016.10.003

Kyle, K., & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly*, *49*(4), 757–786. https://doi.org/10.1002/tesq.194

Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing*, *33*(3), 319–340. https://doi.org/10.1177/0265532215587391

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030–1046. https://doi.org/10.3758/s13428–017–0924–4

Lagrou, E., Hartsuiker, R. J., & Duyck, W. (2011). Knowledge of a second language influences auditory word recognition in the native language. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(4), 952–965. https://doi.org/10.1037/a0023217

Landauer, T. K. (2007). LSA as a theory of meaning. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 3–34). Psychology Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033–295X.104.2.211

Landauer, T. K., Foltz, P., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284.

Lane, H., Howard, C., & Hapke, M. (2019). *Natural Language Processing in action: Understanding, analyzing, and generating text with python*. Manning Publications.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322. https://doi.org/10.1093/applin/16.3.307

Laufer, B., & Shmueli, K. (1997). Memorizing new words: Does teaching have anything to do with it? *RELC Journal*, *28*(1), 89–108. https://doi.org/10.1177/003368829702800106

Laufer, Batia. (1992). How much lexis is necessary for reading comprehension? In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and Applied Linguistics* (pp. 126–132). Palgrave Macmillan UK. https://doi.org/10.1007/978–1–349–12396–4_12

Laufer, Batia, & Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics, 16*(3), 307–322. https://doi.org/10.1093/applin/16.3.307

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods, 44*, 325–343.

Lemhöfer, K., & Dijkstra, T. (2004). Recognizing cognates and interlingual homographs: Effects of code similarity in language-specific and generalized lexical decision. *Memory & Cognition*, *32*(4), 533–550. https://doi.org/10.3758/BF03195845

Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R. H., Grainger, J., & Zwitserlood, P. (2008). Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 12–31. https://doi.org/10.1037/0278–7393.34.1.12

Leow, R. P., Grey, S., Marijuan, S., & Moorman, C. (2014). Concurrent data elicitation procedures, processes, and the early stages of L2 learning: A critical overview. *Second Language Research*, *30*(2), 111–127. https://doi.org/10.1177/0267658313511979

Ling, C. Y., & Braine, G. (2007). The attitudes of university students towards non-native speakers English teachers in Hong Kong. *RELC Journal*, *38*(3), 257–277. https://doi.org/10.1177/0033688207085847

Loper, E., & Bird, S. (2002). NLTK: The Natural Language Toolkit. P*roceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, Philadelphia.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208. https://doi.org/10.3758/BF03204766

MacKenzie, I. (2018). *Language contact and the future of English*. Routledge.

Mainz, N., Shao, Z., Brysbaert, M., & Meyer, A. S. (2017). Vocabulary knowledge predicts lexical processing: Evidence from a group of participants with diverse educational backgrounds. *Frontiers in Psychology*, *8*, 1-14.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Mauranen, A., Hynninen, N., & Ranta, E. (2010). English as an academic lingua franca: The ELFA project. *English for Specific Purposes*, *29*(3), 183-190.

Mazgutova, D., & Kormos, J. (2015). Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of Second Language Writing*, *29*, 3–15. https://doi.org/10.1016/j.jslw.2015.06.004

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, *44*(3), 295–322.

McKinney, W. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, *445*, 51–56.

McNamara, D. S., Allen, L. K., Crossley, S. A., Dascălu, M., & Perret, C. A. (2017). Natural language processing and learning analytics. In C. Lang, G. Siemens, A. Wise, & D.

Gašević (Eds.), *Handbook of Learning Analytics* (First, pp. 93–104). Society for Learning Analytics Research (SoLAR). https://doi.org/10.18608/hla17

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, *23*, 35–59. https://doi.org/10.1016/j.asw.2014.09.002

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, *16*(3), 5–19.

Menn, L., & Dronkers, N. (2017). *Psycholinguistics: Introduction and applications*. Plural Publishing.

Meurers, D. (2013). Natural language processing and language learning. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–13). Blackwell Publishing Ltd.

Meurers, D., & Dickinson, M. (2017). Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning*, *67*(S1), 66–95. https://doi.org/10.1111/lang.12233

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceeding of the International Conference on Learning Representations, Workshop Track*, 1–12.

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751. https://www.aclweb.org/anthology/N13–1090

Monaghan, P., Chang, Y.-N., Welbourne, S., & Brysbaert, M. (2017). Exploring the relations between word frequency, language exposure, and bilingualism in a computational model

of reading. *Journal of Memory and Language*, *93*, 1–21. https://doi.org/10.1016/j.jml.2016.08.003

Monteiro, K. R., Crossley, S. A., & Kyle, K. (2020). In search of new benchmarks: Using L2 lexical frequency and contextual diversity indices to assess second language writing. *Applied Linguistics*, *41*(2), 280–300. https://doi.org/10.1093/applin/amy056

Monteiro, K. R., & Kim, Y. (in press). The effect of input characteristics and individual differences on L2 comprehension of authentic and modified listening tasks. *System*.

Morrison, C. M., & Ellis, A. W. (2000). Real age of acquisition effects in word naming and lexical decision. *British Journal of Psychology*, *91*(2), 167–180. https://doi.org/10.1348/000712600161763

Morrison, C. M., Hirsh, K. W., Chappell, T., & Ellis, A. W. (2002). Age and age of acquisition: An evaluation of the cumulative frequency hypothesis. *European Journal of Cognitive Psychology*, *14*(4), 435–459. https://doi.org/10.1080/09541440143000159

Muncer, S. J., Knight, D., & Adams, J. W. (2014). Bigram Frequency, Number of Syllables and Morphemes and Their Effects on Lexical Decision and Word Naming. *Journal of Psycholinguistic Research*, *43*(3), 241–254. https://doi.org/10.1007/s10936–013–9252–8

Murakami, A. (2016). Modeling systematicity and individuality in nonlinear second language development: The case of English grammatical morphemes. *Language Learning*, *4*(66), 834–871.

Naismith, B., Han, N.-R., Juffs, A., Hill, B. L., & Zheng, D. (2018). Accurate measurement of lexical sophistication with reference to ESL learner data. *Conference Paper in Educational Data Mining*.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142.

Nelson, D. L., & Friedrich, M. A. (1980). Encoding and cuing sounds and senses. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(6), 717–731. https://doi.org/10.1037/0278–7393.6.6.717

O'Donnell, M. B., 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*(1), 83–108. https://doi.org/10.1075/ijcl.18.1.07odo

Ortega, L. (2016). Multi-competence in second language acquisition: Inroads into the mainstream? In V. J. Cook & L. Wei (Eds.), *The Cambridge Handbook of Linguistic Multi-competence* (pp. 50–76). Cambridge University Press.

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). Longman.

Palfreyman, D., & Karaki. (2019). Lexical sophistication across languages: A preliminary study of undergraduate writing in Arabic (L1) and English (L2). *International Journal of Bilingual Education and Bilingualism*, *22*(8), 992–1015.

Pasquarella, A., Chen, X., Gottardo, A., & Geva, E. (2015). Cross-language transfer of word reading accuracy and word reading fluency in Spanish-English and Chinese-English bilinguals: Script-universal and script-specific processes. *Journal of Educational Psychology*, *107*(1), 96–110. https://doi.org/10.1037/a0036966

Perani, D., & Abutalebi, J. (2005). The neural basis of first and second language processing. *Current Opinion in Neurobiology, 15*(2), 202–206. https://doi.org/10.1016/j.conb.2005.03.007

Perea, M., & Gotor, A. (1997). Associative and semantic priming effects occur at very short stimulus-onset asynchronies in lexical decision and naming. *Cognition, 62*(2), 223–240. https://doi.org/10.1016/S0010–0277(96)00782–2

Perfetti, C. (1969). Lexical density and phrase structure depth as variables in sentence retention. *Journal of Verbal Learning and Verbal Behavior, 8*, 719–724.

Pierrehumbert, J. B. (2012). The dynamic lexicon. *Handbook of Laboratory Phonology*, 173–183.

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D. (2017). *Package "nlme." Linear and nonlinear mixed effects models, version, 3.*

Porte, G. (2012). Concluding remarks: The way forward. In *Replication Research in Applied Linguistics* (pp. 268–274). Cambridge University Press.

Portocarrero, J. S., Burright, R. G., & Donovick, P. J. (2007). Vocabulary and verbal fluency of bilingual and monolingual college students. *Archives of Clinical Neuropsychology*, *22*(3), 415–422. https://doi.org/10.1016/j.acn.2007.01.015

Qian, D. D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing, 21*(1), 28–52.

Oliphant, T. E. (2006). *A guide to NumPy* (Vol. 1). Trelgol Publishing USA.

Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, *18*(1), 25–39. https://doi.org/10.1016/j.asw.2012.10.004

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*(3), 355–371. https://doi.org/10.1177/026553229301000308

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *In Proceedings of the Lrec 2010 Workshop on New Challenges for Nlp Frameworks*, 45–50.

Riazi, A. M. (2016). Comparing writing performance in TOEFL-iBT and academic assignments: An exploration of textual features. *Assessing Writing*, *28*, 15–27. https://doi.org/10.1016/j.asw.2016.02.001

Richards, L. G. (1976). Concreteness as a variable in word recognition. *The American Journal of Psychology*, *89*(4), 707–718. https://doi.org/10.2307/1421468

Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, *3*(2), 303–345. https://doi.org/10.1111/j.1756–8765.2010.01111.x

Römer, U. (2009a). Corpus research and practice: What help do teachers need and what can we offer? In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 83–98). John Benjamins Publishing.

Römer, U. (2009b). English in academia: Does nativeness matter? *Anglistik: International Journal of English Studies*, *2*(20), 89–100.

Römer, U. (2016). Teaming up and mixing methods: Collaborative and cross-disciplinary work in corpus research on phraseology. *Corpora*, *11*(1), 113–129. https://doi.org/10.3366/cor.2016.0087

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, *89*(1), 60–94.

Schnoebelen, T., & Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research. *Psihologija*, *43*(4), 441–464.

Scott, S. (1997). PC analysis of key words—And key key words. *System*, *25*(2), 233–245.

Segui, J., & Grainger, J. (1990). Priming word recognition with orthographic neighbors: Effects of relative prime-target frequency. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(1), 65–76.

Shibahara, N., Zorzi, M., Hill, M., Wydell, T., & Butterworth, B. (2003). Semantic effects in word naming: Evidence from English and Japanese Kanji. *The Quarterly Journal of Experimental Psychology*, *56*(2), 263–286.

Shin, Y. K., Cortes, V., & Yoo, I. W. (2018). Using lexical bundles as a tool to analyze definite article use in L2 academic writing: An exploratory study. *Journal of Second Language Writing*, *39*, 29–41. https://doi.org/10.1016/j.jslw.2017.09.004

Shin, Y. K., & Kim, Y. (2017). Using lexical bundles to teach articles to L2 English learners of different proficiencies. *System, 69*, 79–91. https://doi.org/10.1016/j.system.2017.08.002

Sinclair, J. M. (2008). The phrase, the whole phrase, and nothing but the phrase. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 407–410). Amsterdam: John Benjamins.

Skalicky, S., Crossley, S., & Berger, C. (in press). Predictors of second language English lexical recognition: Further insights from a large database of second language lexical decision times. *The Mental Lexicon*.

Skalicky, Stephen, Crossley, S. A., McNamara, D. S., & Muldner, K. (2019). Measuring Creative Ability in Spoken Bilingual Text: The Role of Language Proficiency and Linguistic Features. *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, 1056–1062.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*(1), 155–167.

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes*, *12*(3), 214–225. https://doi.org/10.1016/j.jeap.2013.05.002

Steinhauer, K. (2014). Event-related Potentials (ERPs) in second language research: A brief introduction to the technique, a selected review, and an invitation to reconsider critical periods in L2. *Applied Linguistics*, *35*(4), 393–417. https://doi.org/10.1093/applin/amu028

Ulate, N. (2014). Notions of non-native teachers in Costa Rican language schools. *MexTESOL Journal*, *1*(38), 1–15.

Vaid, J., & Meuter, R. (2017). Language without borders: Reframing the study of the bilingual mental lexicon. In M. Libben, M. Goral, & G. Libben (Eds.), *Bilingualism: A framework for understanding the mental lexicon* (pp. 7–26). John Benjamins Publishing Company.

van Rossum, G. (1995). *Python tutorial. Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI).*

Van Wijnendaele, I., & Brysbaert, M. (2002). Visual word recognition in bilinguals: Phonological priming from the second to the first language. *Journal of Experimental*

Psychology: Human Perception and Performance, *28*(3), 616–627. https://doi.org/10.1037/0096–1523.28.3.616

Vanlangendonck, F., Peeters, D., Rueschemeyer, S.-A., & Dijkstra, T. (2019). Mixing the stimulus list in bilingual lexical decision turns cognate facilitation effects into mirrored inhibition effects. *Bilingualism: Language and Cognition*, 1–9. https://doi.org/10.1017/S1366728919000531

Verkoeijen, P. P. J. L., Rikers, R. M. J. P., & Schmidt, H. G. (2004). Detrimental influence of contextual change on spacing effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 796–800. https://doi.org/10.1037/0278–7393.30.4.796

West, M. (1953). *General service list of English words*. Longman.

Wilks, C., & Meara, P. (2002). Untangling word webs: Graph theory and the notion of density in second language word association networks. *Second Language Research*, *18*(4), 303–324. https://doi.org/10.1191/0267658302sr203oa

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 53–79. https://doi.org/10.1037/a0024177

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, *31*(2), 236–259. https://doi.org/10.1093/applin/amp024

Zareva, A. (2007). Structure of the second language mental lexicon: How does it compare to native speakers' lexical organization? *Second Language Research*, *23*(2), 123–153. https://doi.org/10.1177/0267658307076543

Zhou, H., Chen, B., Yang, M., & Dunlap, S. (2010). Language nonselective access to phonological representations: Evidence from Chinese–English bilinguals. *The Quarterly Journal of Experimental Psychology*, *63*(10), 2051–2066. https://doi.org/10.1080/17470211003718705

APPENDICES

Appendix A: Correlations between EF-CAMDAT Indices and COCA Fiction

	COCA Fiction Indices – Log Transformed									
	AW[a] Freq	***AW Range***	***CW[a] Freq***	***CW Ran-ge***	***FW[a] Freq***	***FW Ran-ge***	***Bi-gram Freq***	***Bi-gram Ran-ge***	***Tri-gram Freq***	***Tri-gram Ran-ge***
EF-CAMDAT AW Freq Log	0.63	-	-	-	-	-	-	-	-	-
EF-CAMDAT AW Range Log	-	0.64	-	-	-	-	-	-	-	-
EF-CAMDAT CW Freq Log	-	-	0.61	-	-	-	-	-	-	-
EF-CAMDAT CW Range Log	-	-	-	0.62	-	-	-	-	-	-
EF-CAMDAT FW Freq Log	-	-	-	-	0.79	-	-	-	-	-
EF-CAMDAT FW Range Log	-	-	-	-	-	0.69	-	-	-	-
EF-CAMDAT Bigrams Freq Log	-	-	-	-	-	-	0.52	-	-	-
EF-CAMDAT Bigrams Range Log	-	-	-	-	-	-	-	0.53	-	-
EF-CAMDAT Trigrams Freq Log	-	-	-	-	-	-	-	-	0.38	-
EF-CAMDAT Trigrams Range Log	-	-	-	-	-	-	-	-	-	0.38

[a] AW = All Words; CW = Content Words; FW = Function Words;

Appendix B: Distribution of Languages for the TOEFL iBT Public Use Dataset

Language	***Number of speakers***	***Language***	***Number of speakers***
Chinese	83	English	3
Korean	56	Farsi	3
Spanish	52	Kannada	3
Japanese	50	Malayalam	3
Arabic	30	Swedish	3
German	26	Urdu	3
French	23	Mongolian	2
Hindi	13	Albania	2
Portuguese	10	Nepali	2

Russian	10	Afrikaans	1
Turkish	9	Akan	1
Telugu	7	Cebuano	1
Tagalog	7	Finnish	1
Gujarati	6	Hebrew	1
Indian	6	Javanese	1
Romanian	6	Khmer	1
Thai	6	Konkani	1
Vietnamese	6	Lithuanian	1
Bengali	5	Macedonian	1
Italian	5	Mende	1
Tamil	5	Norwegian	1
Marthi	4	Polish	1
Greek	4	Somali	1
Yoruba	4	Turkmen	1
Bulgarian	3	Ukrainian	1
Dutch	3	Undefined	1

Appendix C: EF-CAMDAT and COCA Fiction Model Comparisons Statistics

EF-CAMDAT Independent Model Comparisons

		Model Description		***Test Against Prior Model***	
Model	***Fixed Effects***	***Random Effects***	***AIC***	***Statistic***	***p***
1	None	language	1219.7		
2	Model 1 + Age	language	1221.6	X^2 (1) = 0.096	0.75
3	Model 1 + Gender (vs Model 2 refitted to 416 observations, AIC 1068.5)	language	1070.5	X^2 (1) = 0.004	0.95
4	Model 1 + Topic	language	1219.7	X^2 (1) = 2.048	0.15
5	Model 1 + EF-CAMDAT Range of Lemma bigrams Log	language	1173.3	X^2 (1) = 48.196	< .005
6	Model 5 + EF-CAMDAT Frequency of Content Lemmas Log	language	1173.4	X^2 (1) = 2.147	0.14
7	Model 5 + EF-CAMDAT Frequency of Function Lemmas Log	language	1165.2	X^2 (1) = 10.362	< .005

EF-CAMDAT Integrated Model Comparisons

	Model Description			Test Against Prior Model	
Model	***Fixed-Effects***	***Random-Effects***	***AIC***	***Statistic***	***p***
1	None	language	1556.2		
2	Model 1 + Age	language	1558.1	X^2 (1) = 0.038	0.84
3	Model 1 + Gender (vs Model 2 refitted to 416 observations, AIC 1356.3)	language	1358.3	X^2 (1) = 0.001	0.97
4	Model 1 + Topic	language	1558.1	X^2 (1) = 0.064	0.80
5	Model 1 + EF-CAMDAT Frequency of Lemma trigrams Log	language	1537.2	X^2 (1) = 20.951	< .005
6	Model 5 + EF-CAMDAT Range of All Lemmas Log	language	1538.4	X2 (1) = 0.767	0.38
7	Model 5 + EF-CAMDAT Frequency of Lemma bigrams Log	language	1538.6	X^2 (1) = 0.601	0.44
8	Model 5 + EF-CAMDAT Range of Function Lemmas Log	language	1556.7	X^2 (1) = 3.344	0.07

COCA Fiction Independent Model Comparisons

	Model Description			Test Against Prior Model	
Model	***Fixed Effects***	***Random Effects***	***AIC***	***Statistic***	***p***
1	None	language	1219.7		
2	Model 1 + COCA Fiction Range - Content Lemmas Log	language	1151.4	X^2(1) = 70.277	<.005
3	Model 2 + COCA Fiction Lemma bigrams Frequency Log	language	1153.2	X^2 (1) = 0.272	0.60
4	Model 2 + COCA Fiction Lemma trigrams Frequency Log	language	1153.4	X^2 (1) = 0.009	0.92

COCA Fiction Integrated Model Comparisons

Model Description				Test Against Prior Model	
Model	***Fixed Effects***	***Random Effects***	***AIC***	***Statistic***	***p***
1	None	language	1556.2		
2	Model 1 + COCA Fiction Lemma bigrams Frequency Log	language	1551.5	X^2 (1) = 6.61	0.01
3	Model 2 + COCA Fiction Lemma trigrams Frequency Log	language	1552.2	X^2 (1) = 1.291	0.26

Combined Independent Model Comparisons

Model	***Fixed Effects***	***Random Effects***	***AIC***	***Statistic***	***p***
1	None	language	1219.7		
2	Model 1 + COCA Fiction Range - Content lemmas	language	1151.4	X^2 (1) = 70.277	<.005
3	Model 2 + EF-CAMDAT Range - Lemma bigrams log	language	1148.4	X^2 (1) = 5.055	0.02
4	Model 3 + COCA Fiction Frequency - Lemma trigrams log	language	1150.0	X^2 (1) = 0.354	0.55
5	Model 3 + EF-CAMDAT Frequency - Function lemmas log	language	1150.0	X^2 (1) = 0.332	0.56

Combined Integrated Model Comparisons

Model Description				Test Against Prior Model	
Model	***Fixed Effects***	***Random Effects***	***AIC***	***Statistic***	***p***
1	None	language	1556.2		
2	Model 1 + EF-CAMDAT Range - Lemma bigrams log	language	1548.3	X^2 (1) = 9.795	<.005
3	Model 2 + EF-CAMDAT Range - All lemmas log	language	1550.2	X^2 (1) = 0.102	0.75
4	Model 2 + COCA Fiction Frequency - Lemma trigrams log	language	1548.8	X^2 (1) = 1.545	0.21

Appendix D: Individual Output with Frequency and Range Indices – Independent Task

Independent Task - Score 5 – Topic: career choice

I disagree with the statement that studying according to one's interests is more important that studying with focus on a career.

Never before in history, personal development, happiness, and self-realization have been emphasized as much as today. Also, there have never been as many personal choices and possibilities. While in centuries past, women were destined to take up their roles as mothers and housewives, men often had to take over their fathers' trades. This has fortunately changed over the course of the last century. The planning of one's future profession is now a more autonomous decision. But a draw-back of all this newly found freedom is that more people depend on an external job market to find employment and have a career.

While it may be more satisfying from the student's point of view to study something that meets the student's own interests, future employers have an interest in finding employees that meet the requirements of a job description. It is of no concern for an employer whether an applicant loves the fine arts and is a trained artist if the employer is not looking for an artist.

Also, it is quite common that people change their careers over the course of their lives. The most important initial goal to pursue is financial independence. This is best achieved with receiving training in a field that is in high demand. Then, further down the road, the job market may change. In our times of highly valued continued education, it is not hard to pursue a different degree on nights or weekends, or even through the internet at one's own pace. All this can then be accomplished while enjoying a secure income.

Finally, sometimes a personal interest is an excellent hobby that can serve to balance the work life. Going back to the example of the artist, art as a hobby is in my opinion much more enjoyable as a hobby without the pressure to sell. Should the hobby artist come to fame and start to make a fortune (which is not very likely as we all know), a switch to a full-time art career can be made at that time without the first frugal years.

Independent Task - Score 1 – Topic: cooperation

Today most people think that the abillity to coorporate with other people is most important in everyday's life than in the past. Because of the easy way to cooperate with each other. people now can cooporate with each other so easy.

Cooperating systerm have helped people in their business so much and earn more capital than in the past for example long times ago people use so much more time to cooperate with othe, but now they can use computer fo cooperate more faster.

Significant Scores for the High-scored and Low-scored Essays

Scores	***EF-CAMDAT Range - Lemma Bigrams Log***	***COCA Fiction Range - Content Lemmas***	***EF-CAMDAT - Frequency Function Lemmas Log***
5	-6.606	0.167	4.747
1	-5.947	0.1984	4.246

High-scored essay

Bigrams (types)	*Token count*	*EF-CAMDAT Range - Lemma Bigrams Log*	*Content Words (types)*	*Token count*	*COCA Fiction Range - Content Lemmas*	*Funcion Lemmas*	*Token count*	*EF-CAMDAT - Frequency Function Lemmas Log*
external job	1	-12.4	frugal	1	0.001	the	20	6.1
artist also	1	-11.7	self-realization	1	0.001	be	20	6.0
demand then	1	-11.7	housewife	1	0.003	to	13	5.8
find employment	1	-11.7	hobby	4	0.005	i	1	5.8
further down	1	-11.7	enjoyable	1	0.008	and	6	5.7
initial goal	1	-11.7	applicant	1	0.011	a	21	5.7
interest future	1	-11.7	fame	1	0.011	in	7	5.3
pursue be	1	-11.7	destine	1	0.012	of	9	5.1
the hobby	1	-11.7	happiness	1	0.015	have	6	5.0
a switch	1	-11.3	weekend	1	0.015	my	1	4.8
career never	1	-11.3	trained	1	0.018	for	2	4.8
emphasize as	1	-11.3	fortunately	1	0.019	that	9	4.7
history personal	1	-11.3	full-time	1	0.019	it	4	4.5
income finally	1	-11.3	fortune	1	0.020	with	3	4.4
opinion much	1	-11.3	autonomous	1	0.021	we	1	4.3
satisfy from	1	-11.3	employer	3	0.024	this	4	4.3
secure income	1	-11.3	switch	1	0.028	but	1	4.0
and housewife	1	-11.0	continued	1	0.032	can	3	3.9
artist come	1	-11.0	internet	1	0.033	on	3	3.9
career over	1	-11.0	disagree	1	0.035	very	1	3.8
career while	1	-11.0	pace	1	0.035	as	7	3.8
find employee	1	-11.0	artist	4	0.038	more	5	3.7
market may	1	-11.0	employee	1	0.043	not	3	3.7
"ones interest"	1	-10.8	profession	1	0.049	at	2	3.7
have fortunately	1	-10.8	satisfy	1	0.050	all	3	3.6
"at ones"	1	-10.6	excellent	1	0.051	if	1	3.6
artist if	1	-10.6	employment	1	0.052	or	2	3.5
enjoyable as	1	-10.6	newly	1	0.052	our	1	3.4
to fame	1	-10.6	fine	1	0.053	from	1	3.3
today also	1	-10.5	independence	1	0.053	there	1	3.3
personal choice	1	-10.3	income	1	0.054	then	2	3.0
trade this	1	-10.3	night	1	0.058	most	1	2.9
description it	1	-10.2	road	1	0.059	their	4	2.7
this newly	1	-10.2	secure	1	0.062	should	1	2.7
many personal	1	-10.1	father	1	0.065	up	1	2.7
more autonomous	1	-10.1	sell	1	0.065	now	1	2.7
hobby that	1	-10.0	planning	1	0.066	which	1	2.5
of highly	1	-10.0	accomplish	1	0.068	down	1	2.3

Bigrams (types)	*Token count*	*COCA Fiction Range - Lemma Bigrams Log (NOT SIG.)*	*Content Lemmas (types)*	*Token count*	*EF-CAMDAT Range - Content Lemmas (NOT SIG.)*
a external	1	0.001	frugal	1	-11.0
and possibility	1	0.001	self-realization	1	-10.3
art as	1	0.001	destined	1	-8.8
employer be	1	0.001	autonomous	1	-8.3
goal to	1	0.001	"students"	2	-7.5
hobby be	1	0.001	newly	1	-7.3
in century	1	0.001	fame	1	-7.2
likely as	1	0.001	fortune	1	-7.0
may change	1	0.001	housewife	1	-6.8
of highly	1	0.001	initial	1	-6.7
own interest	1	0.001	"ones"	3	-6.6
own pace	1	0.001	independence	1	-6.6
people change	1	0.001	pursue	2	-6.5
personal interest	1	0.001	full-time	1	-6.4
to fame	1	0.001	external	1	-6.4
weekend or	1	0.001	secure	1	-6.2
also it	1	0.002	pace	1	-6.2
can then	1	0.002	switch	1	-6.1
change over	1	0.002	statement	1	-6.0
fine art	1	0.002	emphasize	1	-6.0
freedom be	1	0.002	applicant	1	-5.9
i disagree	1	0.002	hobby	4	-5.9
job description	1	0.002	accomplish	1	-5.8
last century	1	0.002	likely	1	-5.7
market to	1	0.002	employment	1	-5.7
no concern	1	0.002	description	1	-5.5
requirement of	1	0.002	enjoyable	1	-5.5
statement that	1	0.002	artist	4	-5.4
student 's	2	0.002	profession	1	-5.4
that meet	2	0.002	highly	1	-5.4
the example	1	0.002	happiness	1	-5.3
the planning	1	0.002	freedom	1	-5.3
the requirement	1	0.002	trade	1	-5.2
they career	1	0.002	pressure	1	-5.2
they role	1	0.002	disagree	1	-5.1
view to	1	0.002	whether	1	-5.0
a secure	1	0.003	century	2	-4.9
a trained	1	0.003	demand	1	-4.8

receive train	1	-10.0	mother	1	0.068	before	1	2.1
study accord	1	-10.0	career	4	0.071	no	1	2.1
no concern	1	-9.8	external	1	0.072	own	2	1.9
while enjoy	1	-9.8	trade	1	0.073	over	3	1.5
art as	1	-9.8	freedom	1	0.075	while	3	1.3
"to ones"	1	-9.7	love	1	0.075	through	1	0.8
different degree	1	-9.7	pursue	2	0.077	further	1	0.0
highly value	1	-9.7	enjoy	1	0.082			
high demand	1	-9.6	balance	1	0.083			
man often	1	-9.5	financial	1	0.084			
or weekend	1	-9.5	description	1	0.086			
career can	1	-9.5	opinion	1	0.088			
people depend	1	-9.5	requirement	1	0.091			
be destined	1	-9.4	art	3	0.094			
"ones own"	1	-9.4	market	2	0.096			
as mother	1	-9.4	statement	1	0.100			
future profession	1	-9.4	hard	1	0.104			
own pace	1	-9.4	pressure	1	0.104			
freedom be	1	-9.3	training	1	0.104			
be emphasize	1	-9.2	initial	1	0.106			
destined to	1	-9.2	job	3	0.106			
even through	1	-9.2	emphasize	1	0.109			
fame and	1	-9.2	possibility	1	0.116			
fine art	1	-9.1	choice	1	0.120			
very likely	1	-9.0	quite	1	0.120			
on night	1	-9.0	woman	1	0.124			
can then	1	-9.0	sometimes	1	0.127			
in century	1	-9.0	century	2	0.128			
"of ones"	1	-8.9	depend	1	0.129			
with focus	1	-8.9	highly	1	0.129			
as today	1	-8.9	something	1	0.132			
a fortune	1	-8.8	demand	1	0.134			
personal interest	1	-8.8	decision	1	0.137			
"the students"	2	-8.8	man	1	0.142			
a applicant	1	-8.8	achieve	1	0.146			
a external	1	-8.8	degree	1	0.146			
may change	1	-8.7	personal	3	0.146			
past woman	1	-8.7	today	1	0.148			
statement that	1	-8.7	finally	1	0.151			
personal	1	-8.7	field	1	0.153			
field that	1	-8.7	goal	1	0.153			
the statement	1	-8.7	start	1	0.157			
can serve	1	-8.6	never	2	0.158			
whether a	1	-8.6	education	1	0.160			
financial	1	-8.5	student	2	0.162			
never before	1	-8.5	meet	2	0.167			
own interest	1	-8.5	back	1	0.168			
be financial	1	-8.5	history	1	0.170			
pressure to	1	-8.5	likely	1	0.171			
a secure	1	-8.4	past	1	0.175			
switch to	1	-8.4	serve	1	0.176			
internet at	1	-8.4	common	1	0.177			

also there	1	0.003	road	1	-4.8
as mother	1	0.003	serve	1	-4.7
fame and	1	0.003	fortunately	1	-4.7
in find	1	0.003	art	3	-4.6
on night	1	0.003	employer	3	-4.6
pursue a	1	0.003	satisfy	1	-4.6
training in	1	0.003	value	1	-4.5
very likely	1	0.003	balance	1	-4.5
whether a	1	0.003	role	1	-4.5
a full-time	1	0.004	focus	1	-4.4
a hobby	2	0.004	field	1	-4.3
choice and	1	0.004	choice	1	-4.3
field that	1	0.004	mother	1	-4.2
further down	1	0.004	income	1	-4.2
happiness and	1	0.004	history	1	-4.2
interest be	2	0.004	weekend	1	-4.1
make at	1	0.004	accord	1	-4.1
pressure to	1	0.004	development	1	-4.1
study with	1	0.004	possibility	1	-4.1
the statement	1	0.004	sell	1	-4.0
balance the	1	0.005	depend	1	-4.0
be accomplish	1	0.005	past	1	-4.0
disagree with	1	0.005	requirement	1	-4.0
even through	1	0.005	continue	1	-3.9
role as	1	0.005	personal	3	-3.9
's future	1	0.005	fine	1	-3.8
's interest	1	0.005	excellent	1	-3.8
's point	1	0.005	achieve	1	-3.8
should the	1	0.005	common	1	-3.8
time without	1	0.005	quite	1	-3.8
to balance	1	0.005	view	1	-3.8
be destine	1	0.006	financial	1	-3.8
often have	1	0.006	decision	1	-3.7
that more	1	0.006	often	1	-3.7
a switch	1	0.007	internet	1	-3.7
destine to	1	0.007	goal	1	-3.6
important that	1	0.007	receive	1	-3.6
night or	1	0.007	train	2	-3.6
not hard	1	0.007	night	1	-3.5
serve to	1	0.007	degree	1	-3.4
a interest	1	0.008	education	1	-3.3
concern for	1	0.008	today	1	-3.3
road the	1	0.008	employee	1	-3.3
before in	1	0.009	sometimes	1	-3.1
my opinion	1	0.009	without	2	-3.1
the artist	1	0.009	concern	1	-3.1
this can	1	0.009	enjoy	1	-3.1
a career	2	0.010	career	4	-3.0
change they	1	0.010	woman	1	-3.0
in history	1	0.010	opinion	1	-3.0
sometimes a	1	0.010	market	2	-3.0
the internet	1	0.010	something	1	-3.0

that meet	2	-8.4	course	2	0.181
employment and	1	-8.4	future	2	0.181
more satisfy	1	-8.4	last	1	0.181
hobby be	1	-8.4	receive	1	0.181
of no	1	-8.4	accord	1	0.182
not hard	1	-8.3	best	1	0.184
and possibility	1	-8.3	interest	4	0.184
concern for	1	-8.3	value	1	0.185
night or	1	-8.2	look	1	0.186
weekend or	1	-8.2	role	1	0.192
be accomplish	1	-8.2	further	1	0.195
should the	1	-8.2	concern	1	0.196
in find	1	-8.2	development	1	0.197
profession be	1	-8.2	focus	1	0.200
a hobby	2	-8.1	view	1	0.200
continue education	1	-8.1	life	2	0.204
life go	1	-8.1	now	1	0.213
education it	1	-8.1	go	1	0.215
employer have	1	-8.1	high	1	0.216
road the	1	-8.0	own	2	0.220
make at	1	-7.9	people	2	0.221
study something	1	-7.9	point	1	0.223
serve to	1	-7.9	come	1	0.226
pursue a	1	-7.9	often	1	0.226
in history	1	-7.8	very	1	0.230
while it	1	-7.8	different	1	0.232
employer be	1	-7.8	know	1	0.232
a artist	1	-7.8	example	1	0.233
balance the	1	-7.8	important	2	0.237
role as	1	-7.8	change	3	0.238
their role	1	-7.7	study	3	0.238
a field	1	-7.7	much	2	0.240
to pursue	2	-7.6	then	2	0.246
the pressure	1	-7.6	should	1	0.247
employee that	1	-7.5	even	1	0.250
more enjoyable	1	-7.5	work	1	0.261
that study	2	-7.5	find	3	0.265
before in	1	-7.5	take	2	0.268
to balance	1	-7.5	many	1	0.270
requirement of	1	-7.5	year	1	0.270
market to	1	-7.4	first	1	0.271
degree on	1	-7.4	may	2	0.272
up their	1	-7.4	most	1	0.281
decision but	1	-7.4	time	2	0.281
last century	1	-7.4	make	2	0.284
job market	2	-7.3	can	3	0.289
view to	1	-7.3	also	2	0.290
people change	1	-7.3	more	5	0.294
while in	1	-7.3	not	3	0.298
common that	1	-7.3	as	7	0.300
time without	1	-7.3	have	6	0.300
change over	1	-7.3	be	20	0.301

to pursue	2	0.010	man	1	-2.9
while in	1	0.010	example	1	-2.9
a fortune	1	0.011	hard	1	-2.9
we time	1	0.011	may	2	-2.9
art and	1	0.012	never	2	-2.9
enjoy a	1	0.012	different	1	-2.8
more people	1	0.012	course	2	-2.8
switch to	1	0.012	high	1	-2.7
a artist	1	0.013	back	1	-2.7
a excellent	1	0.013	finally	1	-2.7
never before	1	0.013	meet	2	-2.7
example of	1	0.014	plan	1	-2.6
a personal	1	0.015	come	1	-2.5
the fine	1	0.015	even	1	-2.5
of view	1	0.016	love	1	-2.4
all know	1	0.017	important	2	-2.4
while it	1	0.017	study	3	-2.4
of no	1	0.018	start	1	-2.4
they father	1	0.018	point	1	-2.4
a field	1	0.019	interest	4	-2.4
the pressure	1	0.019	change	3	-2.3
the student	2	0.019	last	1	-2.2
most important	1	0.021	future	2	-2.2
more important	1	0.022	live	1	-2.2
that people	1	0.022	much	2	-2.1
then be	1	0.022	best	1	-2.0
the course	2	0.024	life	1	-2.0
as many	1	0.026	find	3	-2.0
in high	1	0.026	many	1	-1.9
to study	1	0.026	look	1	-1.9
course of	2	0.027	first	1	-1.8
be best	1	0.028	know	1	-1.6
it may	1	0.028	job	3	-1.6
's own	2	0.028	take	2	-1.6
not very	1	0.029	go	1	-1.5
to sell	1	0.029	people	2	-1.4
up they	1	0.031	make	2	-1.4
that can	1	0.032	year	1	-1.3
now a	1	0.034	time	2	-1.3
this have	1	0.034	work	1	-1.2
a more	1	0.037			
take over	1	0.037			
depend on	1	0.039			
meet the	2	0.039			
over they	1	0.041			
without the	2	0.041			
interest in	1	0.042			
love the	1	0.042			
they life	1	0.042			
take up	1	0.044			
know a	1	0.045			
focus on	1	0.046			

quite common	1	-7.3
the example	1	-7.2
sometimes a	1	-7.2
future employer	1	-7.2
interest be	2	-7.1
choice and	1	-7.0
a full-time	1	-7.0
then be	1	-7.0
happiness and	1	-6.9
the fine	1	-6.9
take over	1	-6.8
often have	1	-6.8
a employer	1	-6.8
the artist	1	-6.8
century the	1	-6.7
goal to	1	-6.7
their career	1	-6.7
not look	1	-6.6
job description	1	-6.6
study with	1	-6.6
this have	1	-6.6
plan of	1	-6.5
the requirement	1	-6.5
over their	1	-6.5
also there	1	-6.5
that more	1	-6.5
all know	1	-6.4
our time	1	-6.4
art and	1	-6.4
i disagree	1	-6.3
the employer	1	-6.3
without the	2	-6.3
mother and	1	-6.2
the plan	1	-6.2
be best	1	-6.1
disagree with	1	-6.1
important that	1	-6.1
a train	1	-6.1
also it	1	-6.1
train in	1	-6.0
this can	1	-6.0
course of	2	-6.0
now a	1	-6.0
in high	1	-6.0
achieve with	1	-5.9
enjoy a	1	-5.9
live the	1	-5.9
their live	1	-5.9
work life	1	-5.9
there have	1	-5.8
take up	1	-5.7
as many	1	-5.7

the job	1	0.047
much more	1	0.049
that time	1	0.049
the work	1	0.050
one 's	3	0.050
a job	1	0.052
time of	1	0.052
accord to	1	0.055
may be	1	0.055
or even	1	0.055
be quite	1	0.056
point of	1	0.056
woman be	1	0.057
something that	1	0.058
a different	1	0.059
at one	1	0.060
mother and	1	0.063
there have	1	0.064
be of	1	0.065
of one	1	0.069
much as	1	0.073
we all	1	0.073
to one	1	0.075
and start	1	0.077
all this	2	0.078
as much	1	0.084
can be	1	0.084
in we	1	0.085
hard to	1	0.087
as we	1	0.088
at that	1	0.090
the road	1	0.090
not look	1	0.092
but a	1	0.095
be now	1	0.096
never be	1	0.096
of all	1	0.112
start to	1	0.115
be make	1	0.117
go back	1	0.117
look for	1	0.118
if the	1	0.124
the most	1	0.130
be as	1	0.135
and have	1	0.138
be more	2	0.139
to find	1	0.139
which be	1	0.144
have never	1	0.151
in my	1	0.158
and be	1	0.161
to take	2	0.163

know a	1	-5.5
be of	1	-5.5
much as	1	-5.5
a personal	1	-5.5
meet the	2	-5.5
example of	1	-5.4
be now	1	-5.4
the course	2	-5.4
more important	1	-5.4
change their	1	-5.3
something that	1	-5.3
the road	1	-5.3
never be	1	-5.3
as much	1	-5.3
a interest	1	-5.3
to sell	1	-5.3
a different	1	-5.3
time of	1	-5.2
or even	1	-5.2
love the	1	-5.2
be as	1	-5.2
a more	1	-5.1
but a	1	-5.1
a career	2	-5.1
through the	1	-5.0
of view	1	-5.0
we all	1	-5.0
focus on	1	-5.0
not very	1	-5.0
more people	1	-4.9
at that	1	-4.9
have never	1	-4.8
a excellent	1	-4.8
may be	1	-4.8
that can	1	-4.8
change in	1	-4.7
much more	1	-4.7
and start	1	-4.7
as we	1	-4.6
point of	1	-4.6
of their	1	-4.6
that people	1	-4.6
woman be	1	-4.5
be quite	1	-4.5
that time	1	-4.5
all this	2	-4.4
go back	1	-4.3
the internet	1	-4.3
accord to	1	-4.3
be make	1	-4.3
depend on	1	-4.3
the work	1	-4.3

the last	1	0.168
down the	1	0.180
make a	1	0.180
be that	1	0.186
to make	1	0.188
come to	1	0.190
back to	1	0.196
over the	2	0.202
the first	1	0.203
of they	1	0.212
this be	1	0.213
through the	1	0.213
to a	1	0.216
as a	2	0.217
on a	2	0.220
be in	2	0.228
have to	1	0.241
have a	2	0.244
for a	2	0.247
with the	1	0.248
of a	1	0.251
that be	1	0.257
have be	1	0.258
from the	1	0.261
in a	1	0.265
be not	3	0.271
to the	1	0.279
be a	2	0.280
it be	3	0.284
of the	2	0.285

hard to	1	-4.3
down the	1	-4.2
over the	2	-4.0
come to	1	-4.0
most important	1	-3.9
if the	1	-3.8
start to	1	-3.8
back to	1	-3.8
it may	1	-3.6
the job	1	-3.5
of all	1	-3.5
interest in	1	-3.5
to find	1	-3.5
to study	1	-3.4
which be	1	-3.4
a job	1	-3.4
the last	1	-3.4
my opinion	1	-3.3
in our	1	-3.3
be that	1	-3.2
and have	1	-3.2
and be	1	-3.2
on a	2	-3.2
can be	1	-3.1
look for	1	-3.1
the first	1	-3.1
from the	1	-3.1
of a	1	-3.1
to a	1	-3.1
be more	2	-2.8
make a	1	-2.8
to take	2	-2.8
be in	2	-2.7
as a	2	-2.6
that be	1	-2.5
to make	1	-2.5
this be	1	-2.5
have be	1	-2.3
with the	1	-2.2
be not	3	-2.2
the most	1	-2.2
for a	2	-2.0
to the	1	-2.0
in a	1	-1.9
have to	1	-1.9
in my	1	-1.9
it be	3	-1.5
of the	2	-1.5
have a	2	-1.4
be a	2	-1.1

Low-scored essay

Bigrams	*Token count*	*EF-CAMDAT Range - Lemma Bigram Log*
cooporate with	1	-11.721
with othe	1	-11.721
more capital	1	-11.316
abillity to	1	-11.028
cooperate more	1	-11.028
the abillity	1	-11.028
everyday's life	1	-10.805
today most	1	-9.156
past for	1	-9.119
past because	1	-8.804
now can	1	-8.337
life than	1	-8.32
ago people	1	-8.195
use so	1	-8.045
people now	1	-7.705
to cooperate	2	-7.678
cooperate with	2	-7.669
business so	1	-7.41
more fast	1	-7.273
the easy	1	-7.126
other so	1	-7.021
use computer	1	-6.963
easy way	1	-6.82
have help	1	-6.548
and earn	1	-6.22
people use	1	-6.05
so easy	1	-6.003
than in	2	-6.003
time ago	1	-5.967
now they	1	-5.83
people think	1	-5.819
with each	2	-5.745
earn more	1	-5.577
important in	1	-5.556
their business	1	-5.531
be most	1	-5.484
much and	1	-5.022
much more	1	-4.72
most people	1	-4.547
but now	1	-4.51
with other	1	-4.509
help people	1	-4.475
more time	1	-4.451
can use	1	-4.397
they can	1	-4.375
in their	1	-4.367
other people	2	-4.281
the past	2	-4.201
most important	1	-3.946
each other	2	-3.841
people in	1	-3.827
so much	2	-3.746
long time	1	-3.565
way to	1	-3.557
time to	1	-3.421
people be	1	-3.42
because of	1	-3.27
for example	1	-3.152

Content Words (types)	*Token count*	*COCA Fiction Range - Content Lemmas*
fo	1	0.002001
cooperate	4	0.010264
everyday	1	0.014294
capital	1	0.026779
earn	1	0.046046
example	1	0.04837
faster	1	0.050902
computer	1	0.054236
important	1	0.10724
easy	2	0.119492
business	1	0.128367
today	1	0.135912
past	2	0.182458
help	1	0.20738
most	2	0.218951
use	2	0.225743
people	5	0.228374
life	1	0.232324
much	2	0.240779
long	1	0.24859
way	1	0.260723
can	2	0.262758
other	3	0.263127
more	3	0.264268
think	1	0.268441
now	2	0.268587
time	2	0.274794
so	3	0.28102
have	1	0.29376
be	1	0.299597

Function lemmas	*Token count*	*EF-CAMDAT Frequency - Function Lemmas Log*
the	4	6.117
be	1	6.017
to	3	5.814
and	1	5.673
in	4	5.27
of	1	5.127
have	1	4.97
for	1	4.803
that	1	4.692
with	4	4.426
but	1	3.977
can	2	3.886
more	3	3.736
so	3	3.668
they	1	3.488
because	1	3.418
other	3	3.059
each	2	3.022
most	2	2.942
their	1	2.725
now	2	2.671
than	2	2.408

think that	1	-2.643
that the	1	-2.54
of the	1	-1.477
in the	2	-1.108

Appendix E: Individual Output with Frequency and Range Indices – Integrated Task

Integrated – Score 5 - Topic: fish farming

The claims of the passage was rebutted by the professor. The professor comes up with a counter argument for each of the earlier claims of the passage. Farming which is considered to be harmful for the wild fish in the area around the farms due to the infection spread from the hatcheries to the wild counterparts is questioned by the professor, who argues that the traditional commercial fishing is much more detrimetal to the wild fish than the farms effect on them. He also states that the local commercial fishing had reduced the wild fish density along the shoreline and hence the fishing farms cannot spread the infection in large scale.

Farm fish are fed with growth-inducing chemicals, which affects the human health when consumed. The authors argues that the poultry, pork and beef are also contaminated by the chemicals. He opines that fish has better nutritional value compared to the poultry or other forms of meat. Since we consume the other forms of meet with no complains, he suggests we might as well eat fish. His argument is since all forms of meat have the artificial chemical influence so blaming only the farm fish to be chemically harmful is baseless.

Finally the professor argues the claim that fish farms relates to long-term wastefulness of the process, acoording to him, the fish which are fed to the fish of the farm are the ones which are not edible by the humans. He hints that inedible fish is converted to the edible form on the contrary. According to the professor the fish farming is not at all harmful in anyway to the population in the wild. On the contrary it is in a way helpful to humans.

Integrated – Score 1 - Topic: fish farming

Fish farming has increased of commercial fish production for about 50 years ago. Fish farming is consuming one third of fish demand these days. However, fish farming isn't edible unless that farmers give them some chemical substances that make them edible, thats the first negative of fish farmers. The secnod one is that farm-raised fish may pose a health risk to human consumers in order to produce bigger fish faster. However, there is no negative results of that yet. Although fish has less fat and its good for health due to the amount that it has of protin, fish farming is not because of using pounds of fish meal in order to produce one pound of farmed fish, because of that the amount of protin is decreasing. Fish farmers became endangered from the tradition. Although, fish farmers take care of their fish and use treat them if they have to, fish may spread the diseases easily in their surrounding waters due to the huge number that they swimming at.

Significant Scores for the High-scored and Low-scored Essays

Scores	***EF-CAMDAT Frequency - Lemma Trigrams Log***	***COCA Fiction Frequency – Lemma Bigrams Log***
5	-5.013	0.920
1	-4.013	1.150

High-scored						Low-scored					
Trigrams (types)	*Token count*	*EF-CAMDAT Frequency - Lemma Trigrams Log*	*Bigrams (types)*	*Token count*	*COCA Fiction Frequency - Bigram Lemma Log*	*Trigrams (types)*	*Token count*	*EF-CAMDAT Frequency - Lemma Trigrams Log*	*Bigrams (types)*	*Token count*	*COCA Fiction Frequency - Bigram Lemma Log*
affect the human	1	-7.193	fish which	1	0.186	farm be not	1	-7.193	unless that	1	0.204
beef be conjurer	1	-7.193	complain he	1	0.204	have less fat	1	-7.193	one third	1	0.218
by the chemical	1	-7.193	fish of	1	0.227	health due to	1	-7.193	they fish	1	0.218
farm be not	1	-7.193	all form	1	0.234	less fat and	1	-7.193	substance that	1	0.234
fish to be	1	-7.193	in anyway	1	0.234	number that they	1	-7.193	swim at	1	0.250
of the passage	2	-7.193	fishing be	1	0.237	the first negative	1	-7.193	of commercial	1	0.253
the claim that	1	-7.193	that fish	2	0.250	one pound of	1	-7.193	50 year	1	0.262
the fish of	1	-7.193	eat fish	1	0.271	thats the first	1	-6.787	they surround	1	0.294
the other form	1	-7.193	not spread	1	0.274	have increase of	1	-6.276	easily in	1	0.297
the passage be	1	-7.193	and hence	1	0.280	and its good	1	-6.094	fish have	1	0.332
the professor who	1	-7.193	argument for	1	0.305	water due to	1	-5.94	they swim	1	0.362
which be feed	1	-7.193	of meet	1	0.305	the huge number	1	-5.688	risk to	1	0.390
pork and beef	1	-7.193	he argument	1	0.308	the amount that	1	-5.583	a health	1	0.399
the fish which	1	-7.193	the infection	2	0.308	this day however	1	-5.583	however there	1	0.412
who argue that	1	-7.193	helpful to	1	0.311	because of use	1	-5.321	have increase	1	0.443
and hence the	1	-6.787	passage be	1	0.316	order to produce	2	-5.321	number that	1	0.453
contaminate by	1	-6.787	anyway to	1	0.321	to the huge	1	-5.247	pose a	1	0.468
them he conjurer	1	-6.787	have reduce	1	0.321	that the amount	1	-4.941	that yet	1	0.492
to the fish	1	-6.787	other form	2	0.321	one third of	1	-4.841	meal in	1	0.502
be convert to	1	-6.499	pork and	1	0.329	to the amount	1	-4.75	the tradition	1	0.534
by the professor	1	-6.499	fish have	1	0.332	its good for	1	-4.518	treat they	1	0.580
fish of the	1	-6.499	question by	1	0.350	be not because	1	-4.484	have less	1	0.622
fish which be	1	-6.499	feed to	1	0.359	give them some	1	-4.275	be consume	1	0.625
form on the	1	-6.499	professor	1	0.362	not be-cause of	1	-4.172	they some	1	0.648
or other form	1	-6.499	claim of	2	0.371	result of that	1	-4.124	to fish	1	0.661
by the profe-ssor	1	-6.499	spread from	1	0.381	that make them	1	-4.036	to human	1	0.704
be feed with	1	-6.276	since all	1	0.385	good for health	1	-3.954	the disease	1	0.797
farm which be	1	-6.276	argue the	1	0.422	them if they	1	-3.878	third of	1	0.838
for the wild	1	-6.276	area around	1	0.433	of that the	1	-3.261	of use	1	0.850
conjurer state that	1	-6.276	consume the	1	0.439	one be that	1	-3.09	fat and	1	0.855
claim of the	2	-6.094	argument be	1	0.451	care of their	1	-2.704	the amount	2	0.879
he suggest we	1	-6.094	consider to	1	0.472	that it have	1	-2.384	they if	1	0.887
might as well	1	-6.094	fish to	1	0.497	if they have	1	-2.111	spread the	1	0.888
question by the	1	-5.94	hence the	1	0.501	however there be	1	-1.722	of fish	3	0.890
which affect the	1	-5.94	farm be	1	0.518	the amount of	1	-1.558	to produce	2	0.894
harmful for the	1	-5.94	a counter	1	0.531	because of that	1	-1.415	have of	1	0.940
the farm be	1	-5.806	the claim	2	0.536	due to the	2	-0.867	pound of	2	1.001
his argument be	1	-5.688	the artificial	1	0.539	they have to	1	-0.494	for about	1	1.028
the human health	1	-5.688	be convert	1	0.561	there be no	1	0.186	fish and	1	1.029
in large scale	1	-5.583	the earlier	1	0.585	take care of	1	0.245	result of	1	1.251
of the farm	1	-5.583	the shoreline	1	0.587	in order to	2	1.118	not because	1	1.280
to the wild	2	-5.583	the fishing	1	0.615				due to	2	1.312
be question by	1	-5.488	the chemical	1	0.619				and use	1	1.326
contrary it be	1	-5.488	hint that	1	0.623				the huge	1	1.357
to the professor	1	-5.488	in large	1	0.649				amount of	1	1.452
the claim of	1	-5.488	be question	1	0.655				good for	1	1.520
the author argue	1	-5.401	reduce the	1	0.655				these day	1	1.604
author argue that	1	-5.321	fish be	2	0.668				and its	1	1.630
that the traditional	1	-5.321	form on	1	0.694				give they	1	1.670
all form of	1	-5.178	suggest we	1	0.694				in order	2	1.712
that the local	1	-5.178	affect the	1	0.695				order to	2	1.725
area around the	1	-5.178	to human	1	0.704				make they	1	1.811
the area around	1	-5.113	one which	1	0.710				take care	1	1.885
be harm-ful for	1	-5.052	the	1	0.733				care of	1	1.898
the contrary it	1	-4.995	be since	1	0.794				one be	1	1.905

by the human	1	-4.941
have reduce the	1	-4.941
of the early	1	-4.941
in the wild	1	-4.667
the shoreline and	1	-4.667
other form of	2	-4.484
which be consider	1	-4.275
fish in the	1	-4.222
the one which	1	-4.197
one which be	1	-4.036
popular-tion in the	1	-4.036
to him the	1	-3.994
to the po-pulation	1	-3.974
state that the	1	-3.954
of the process	1	-3.915
for each of	1	-3.86
the po-pulation in	1	-3.825
be not at	1	-3.666
argue that the	2	-3.261
not at all	1	-3.232
each of the	1	-3.09
consider to be	1	-3.042
be con-sider to	1	-2.888
up with a	1	-2.577
in a way	1	-2.443
come up with	1	-2.352
compa-re to the	1	-2.306
which be not	1	-2.237
on the contrary	2	-2.039
be much more	1	-1.482
it be in	1	-1.456
in the area	1	-1.407
be the one	1	-1.037
due to the	1	-0.867
accord to the	1	-0.632
be in a	1	-0.045

he suggest	1	0.805
the	1	0.805
convert to	1	0.827
state that	1	0.828
argue that	2	0.869
spread the	1	0.888
fish in	1	0.892
of meat	2	0.900
the contrary	2	0.915
effect on	1	0.948
relate to	1	0.966
be feed	2	0.990
the passage	2	1.005
have better	1	1.040
claim that	1	1.056
or other	1	1.102
finally the	1	1.174
the professor	5	1.180
meet with	1	1.203
the wild	5	1.219
the author	1	1.234
compare to	1	1.251
for each	1	1.258
since we	1	1.275
due to	1	1.312
be consider	1	1.313
the farm	4	1.334
the fish	3	1.359
we might	1	1.362
he also	1	1.369
the process	1	1.380
might as	1	1.415
the area	1	1.468
form of	3	1.502
the human	2	1.562
the local	1	1.603
much more	1	1.613
each of	1	1.625
not at	1	1.660
accord to	1	1.695
be much	1	1.696
with no	1	1.735
only the	1	1.860
be also	1	1.948
a way	1	1.969
he the	1	2.006
than the	1	2.072
come up	1	2.099
on they	1	2.143
up with	1	2.161
along the	1	2.191
as well	1	2.248
at all	1	2.326
which be	3	2.360
the one	1	2.421
around the	1	2.452
have the	1	2.498
that the	3	2.681
the other	1	2.804
by the	4	2.849
can not	1	2.851
be in	1	2.857
to he	1	2.871
for the	1	3.056
with a	1	3.067
to be	2	3.140
from the	1	3.160
in a	1	3.184
be the	1	3.213

that make	1	1.942
because of	2	1.993
year ago	1	2.100
of that	2	2.214
if they	1	2.238
that they	1	2.268
in they	1	2.310
that it	1	2.358
it have	1	2.519
be that	1	2.603
be no	1	2.667
they have	1	2.669
the first	1	2.675
that the	1	2.681
of they	1	2.796
have to	1	3.042
from the	1	3.160
there be	1	3.239
be not	2	3.393
to the	2	3.449

be not	2	3.393
on the	2	3.447
to the	8	3.449
it be	1	3.635
in the	2	3.652
of the	5	3.675

Appendix F: Frequency and Range Scores for the 100 Words with Higher and Lower RT and Accuracy scores

word	*RT*	*EF-CAMDAT Range - All Words Log*	*COCA Fiction Range - All Words Log*
trek	1303.494	-9.323	-1.700
sine	1294.850	-9.642	-2.705
sermon	1120.656	-11.028	-1.777
lager	1101.459	-10.805	-2.577
stud	1052.056	-9.642	-1.850
lesser	1037.467	-8.586	-1.593
chapel	1033.796	-9.775	-1.584
verse	1031.210	-9.323	-1.627
shrub	1028.813	-12.414	-2.321
suspend	1027.874	-9.156	-2.317
linen	1023.743	-9.370	-1.312
sow	1023.286	-8.110	-2.127
philosopher	1020.770	-9.323	-1.884
mower	1020.563	-11.028	-2.037
gala	1018.577	-9.849	-2.244
attend	1013.967	-4.497	-1.159
paddy	1012.067	-10.805	-2.307
gown	1010.856	-9.706	-1.219
tinker	998.891	-11.028	-2.213
anthem	993.164	-9.849	-2.217
tornado	989.865	-7.555	-1.855
outer	987.880	-8.287	-1.208
porridge	986.962	-10.623	-2.192
curve	986.574	-8.564	-1.154
wag	986.043	nan	-2.260
giggle	984.904	nan	-1.497
adjacent	977.493	-9.470	-1.535
accent	973.232	-5.516	-1.027
gram	971.959	-10.112	-2.260
rarely	970.848	-6.050	-1.002
query	968.172	-9.470	-2.027
fascist	967.052	-11.028	-2.321
cauliflower	966.843	-11.028	-2.561
grocer	965.902	-11.721	-2.236
tramp	965.346	-10.623	-2.013
continent	963.629	-7.162	-1.535
industrial	963.232	-5.819	-1.478
cricket	962.829	-10.112	-2.023
irrelevant	962.815	-9.156	-1.693
loser	962.725	-8.287	-1.674
refer	962.554	-7.087	-1.570
noun	962.533	-10.335	-2.410
spouse	961.907	-8.543	-1.947
campus	961.131	-7.404	-1.390
sick	960.743	-4.580	-0.653
proposition	960.717	-7.780	-1.830
dice	959.164	-10.217	-1.802
instance	958.240	-5.216	-1.177
expression	957.535	-6.787	-0.672

word	*Accuracy*	*EF-CAMDAT Range - All Words Log*	*COCA Fiction Range - All Words Log*
paddy	0.250	-10.805	-2.307
trot	0.263	-11.316	-1.666
mousse	0.316	-8.586	-2.399
muck	0.316	-10.469	-1.725
gala	0.368	-9.849	-2.244
sine	0.368	-9.642	-2.705
sow	0.389	-8.110	-2.127
lager	0.444	-10.805	-2.577
mare	0.450	-9.323	-1.767
posh	0.450	-9.642	-2.145
treble	0.455	-10.623	-2.604
trek	0.474	-9.323	-1.700
bog	0.500	-10.217	-2.178
loom	0.500	-11.721	-1.925
sod	0.524	-11.721	-2.100
mince	0.550	-11.721	-2.529
basin	0.556	-10.217	-1.571
hinge	0.556	-11.721	-2.030
grocer	0.579	-11.721	-2.236
barge	0.588	-10.623	-1.929
nuisance	0.588	-10.112	-1.812
apex	0.600	-10.805	-2.202
gin	0.600	-10.623	-1.531
wary	0.611	-7.996	-1.378
con	0.625	-8.256	-1.565
horrid	0.625	-10.805	-1.857
owe	0.632	-7.273	-1.289
tart	0.632	-6.921	-1.978
wit	0.632	-8.225	-1.493
atom	0.636	-10.217	-2.095
hum	0.636	-9.156	-1.268
dwarf	0.647	-11.721	-1.854
hearth	0.647	-8.586	-1.680
ale	0.650	-8.630	-1.817
mason	0.650	-8.888	-1.821
semi	0.650	-9.419	-2.209
twig	0.650	-11.316	-1.857
axis	0.667	-9.524	-1.941
dam	0.667	-7.893	-1.732
digger	0.667	-11.316	-2.410
gown	0.667	-9.706	-1.219
inner	0.667	-7.116	-1.041
linen	0.667	-9.370	-1.312
mend	0.667	-9.849	-1.939
swan	0.667	-8.751	-1.849
triumph	0.667	-9.581	-1.362
dale	0.682	-11.316	-2.142
assimilate	0.684	-9.524	-2.486
elm	0.684	-6.241	-1.982

knit	956.908	-9.930	-1.600
pin	956.300	-3.232	-1.256
circumstance	954.857	-7.048	-1.734
mayor	954.380	-6.780	-1.531
payment	950.756	-4.485	-1.485
drag	949.374	-9.013	-1.021
hinge	948.702	-11.721	-2.030
bureau	948.125	-8.523	-1.473
dam	947.949	-7.893	-1.732
erect	947.914	-10.335	-1.504
cliff	947.896	-8.677	-1.344
stump	947.020	-11.721	-1.603
suffer	943.445	-5.201	-1.192
stallion	943.257	-11.721	-1.993
mend	941.000	-9.849	-1.939
manufacture	940.635	-7.167	-2.081
distinguished	940.253	-8.543	-1.536
border	939.792	-7.487	-1.207
poisonous	939.234	-9.930	-1.718
particular	939.000	-5.422	-0.760
nuisance	936.488	-10.112	-1.812
frightened	936.254	-6.998	-0.919
intake	935.744	-9.581	-1.867
tanker	934.218	-11.028	-2.416
frightening	932.616	-7.473	-1.317
prompt	930.428	-7.255	-1.978
sword	929.791	-6.574	-1.320
socialism	929.069	-9.323	-2.428
lieutenant	928.180	-9.581	-1.482
tart	928.068	-6.921	-1.978
blonde	928.020	-8.389	-1.188
loom	927.302	-11.721	-1.925
con	926.895	-8.256	-1.565
pillar	925.844	-9.930	-1.772
axis	923.788	-9.524	-1.941
wardrobe	923.451	-8.166	-1.542
hum	922.762	-9.156	-1.268
unnecessary	922.684	-6.408	-1.562
vase	922.667	-8.831	-1.583
riot	922.255	-9.775	-1.676
wander	921.684	-9.156	-1.274
different	921.386	-2.839	-0.377
revise	921.309	-8.271	-2.351
excitement	921.071	-7.996	-0.960
census	920.633	-10.335	-2.382
width	920.615	-6.379	-1.770
frequent	919.832	-6.609	-1.479
dense	919.823	-9.419	-1.262
back	919.756	-2.722	-0.060
burglary	919.624	-8.726	-2.248
colleague	919.358	-5.215	-1.535
city	607.948	-3.154	-0.441
nation	607.568	-6.671	-1.274
hire	607.089	-4.883	-1.272
book	606.925	-4.368	-0.526
free	606.895	-3.353	-0.477
floor	606.500	-4.909	-0.335
agent	606.171	-5.182	-1.142
stock	605.529	-6.490	-1.093
broken	604.632	-5.362	-0.560
village	604.397	-5.399	-0.942
pen	603.938	-5.598	-0.996
grass	603.915	-7.072	-0.722
rather	603.728	-4.691	-0.496
go	603.564	-2.157	-0.111
seven	603.328	-5.213	-0.610
command	603.257	-7.244	-1.058
play	603.088	-3.147	-0.476
cow	603.064	-8.320	-1.276

filthy	0.684	-9.323	-1.282
lily	0.684	-7.509	-1.686
rifle	0.684	-10.017	-1.302
saucer	0.684	-10.469	-1.722
cod	0.700	-9.419	-1.941
gallop	0.700	-11.721	-1.931
tit	0.700	-11.028	-2.124
diminish	0.706	-8.859	-1.975
mop	0.706	-10.217	-1.625
peg	0.706	-10.469	-1.862
rake	0.706	-10.623	-1.812
reel	0.706	-11.028	-1.777
dart	0.714	-10.217	-1.657
omit	0.714	-9.775	-2.604
bulb	0.722	-10.017	-1.506
concede	0.722	-10.335	-1.975
dread	0.722	-10.469	-1.278
feast	0.722	-8.483	-1.440
monarchy	0.722	-10.623	-2.577
mower	0.722	-11.028	-2.037
starve	0.722	-9.581	-1.737
width	0.722	-6.379	-1.770
reed	0.727	-9.236	-1.754
dice	0.733	-10.217	-1.802
comrade	0.737	-10.335	-1.943
dilute	0.737	-11.721	-2.595
foam	0.737	-9.775	-1.447
hay	0.737	-9.419	-1.428
hip	0.737	-8.407	-1.062
ivy	0.737	-8.831	-1.592
lecturer	0.737	-7.850	-2.244
ton	0.737	-8.425	-1.653
varnish	0.737	-11.721	-2.248
vow	0.737	-10.217	-1.792
bureau	0.750	-8.523	-1.473
fist	0.750	-6.783	-0.898
pigeon	0.750	-11.028	-1.804
rub	0.750	-8.354	-1.266
shore	0.750	-7.371	-1.038
vital	0.750	-7.048	-1.410
yacht	0.750	-8.831	-1.937
bias	0.762	-9.849	-2.276
holly	0.762	-9.279	-1.959
asylum	0.765	-9.930	-2.020
fare	0.765	-8.097	-1.632
iron	0.765	-6.506	-0.894
lieutenant	0.765	-9.581	-1.482
thump	0.765	-11.316	-1.487
dust	0.773	-8.195	-0.721
ion	0.773	-9.706	-2.303
bra	0.778	-10.469	-1.428
question	1.000	-4.413	-0.471
break	1.000	-4.895	-0.550
factory	1.000	-5.767	-1.330
clothes	1.000	-4.699	-0.495
probability	1.000	-8.304	-1.980
brilliant	1.000	-5.451	-0.969
emotion	1.000	-6.659	-1.126
competition	1.000	-5.427	-1.380
jump	1.000	-5.145	-0.882
get	1.000	-1.699	-0.107
infinity	1.000	-9.370	-1.855
fate	1.000	-7.871	-1.009
little	1.000	-2.411	-0.123
human	1.000	-4.122	-0.580
treasure	1.000	-7.532	-1.366
blank	1.000	-8.152	-0.983
distribute	1.000	-7.871	-2.133
construction	1.000	-5.417	-1.189

angry	603.027	-5.231	-0.660
positive	602.702	-4.015	-1.260
knock	602.557	-3.939	-0.937
calendar	602.429	-7.661	-1.478
good	602.116	-1.435	-0.129
wipe	601.795	-9.119	-1.171
opera	601.743	-5.441	-1.482
chicken	601.525	-5.959	-0.963
box	600.813	-5.579	-0.597
pause	600.174	-8.152	-0.891
sky	600.147	-5.145	-0.476
wine	600.062	-4.600	-0.852
plane	598.751	-5.148	-1.000
hospital	598.627	-4.696	-0.762
thinking	598.362	-3.713	-0.360
where	598.078	-2.733	-0.094
advice	598.027	-4.138	-1.004
son	597.961	-4.296	-0.511
physical	597.882	-4.782	-0.894
use	597.879	-2.245	-0.389
local	597.861	-4.052	-0.724
index	597.740	-7.861	-1.333
office	597.177	-3.150	-0.519
hood	597.031	-9.279	-1.166
gay	596.706	-7.404	-1.379
root	596.647	-7.770	-1.270
broker	595.678	-7.547	-2.025
camera	595.575	-5.970	-0.964
imagination	595.147	-6.981	-1.053
beer	595.115	-5.602	-0.831
energy	594.936	-4.193	-0.862
drive	594.858	-4.638	-0.566
year	594.764	-2.056	-0.315
tooth	594.669	-9.156	-1.393
dress	594.342	-5.517	-0.630
six	593.669	-4.431	-0.415
white	591.258	-4.236	-0.248
fish	590.945	-5.495	-0.772
buy	590.893	-2.995	-0.612
junior	590.294	-6.386	-1.145
idea	589.910	-3.830	-0.386
lunch	589.848	-4.263	-0.727
signal	588.948	-7.131	-1.082
shampoo	588.005	-10.017	-1.777
happy	586.941	-2.605	-0.464
stream	586.838	-7.147	-0.987
pay	586.778	-2.772	-0.553
onion	586.162	-7.678	-1.708
conclusion	585.476	-4.828	-1.313
provide	585.095	-4.390	-1.088
pick	584.730	-5.541	-0.564
listen	583.600	-3.820	-0.535
love	582.612	-2.581	-0.306
black	582.090	-5.095	-0.253
bath	581.175	-7.082	-1.118
bowl	580.471	-7.586	-0.897
map	580.203	-5.370	-1.057
computer	579.268	-3.074	-0.918
beginning	578.632	-4.170	-0.590
public	578.421	-3.764	-0.721
part	578.008	-3.436	-0.305
expect	577.005	-4.944	-0.720
music	576.542	-3.533	-0.632
couch	575.305	-6.151	-0.905
fact	574.174	-3.517	-0.396
fast	573.158	-4.086	-0.532
right	572.673	-3.108	-0.118
express	570.646	-5.137	-1.248
balloon	568.568	-6.398	-1.518

personal	1.000	-3.861	-0.727
desk	1.000	-6.011	-0.639
space	1.000	-4.760	-0.562
stall	1.000	-9.849	-1.416
recall	1.000	-7.315	-0.972
crossing	1.000	-7.547	-1.059
may	1.000	-2.886	-0.407
police	1.000	-4.479	-0.731
progress	1.000	-4.940	-1.043
luxury	1.000	-6.055	-1.372
socialism	1.000	-9.323	-2.428
fasten	1.000	-9.156	-2.049
prevent	1.000	-5.429	-1.244
come	1.000	-3.138	-0.117
random	1.000	-8.180	-1.158
butterfly	1.000	-9.642	-1.551
philosophy	1.000	-7.126	-1.419
negative	1.000	-5.074	-1.418
perceive	1.000	-7.893	-1.812
begin	1.000	-4.691	-0.644
collapse	1.000	-7.661	-1.333
war	1.000	-5.265	-0.640
morality	1.000	-8.804	-1.921
again	1.000	-3.131	-0.124
skin	1.000	-6.688	-0.432
council	1.000	-5.347	-1.428
page	1.000	-5.888	-0.853
fight	1.000	-5.087	-0.692
flood	1.000	-5.421	-1.318
place	1.000	-2.838	-0.191
stuff	1.000	-5.570	-0.612
fellowship	1.000	-8.354	-1.949
chase	1.000	-7.267	-1.174
abroad	1.000	-4.272	-1.650
noisy	1.000	-6.893	-1.320
client	1.000	-3.914	-1.332
comic	1.000	-8.032	-1.468
edge	1.000	-6.885	-0.488
direction	1.000	-5.736	-0.635
tonight	1.000	-6.745	-0.639
automatic	1.000	-7.391	-1.353
attack	1.000	-6.021	-0.878
advertising	1.000	-5.353	-1.523
capture	1.000	-6.854	-1.411
argue	1.000	-6.787	-1.099
advertisement	1.000	-5.191	-1.988
conclusion	1.000	-4.828	-1.313
bring	1.000	-3.912	-0.441
scrape	1.000	nan	-1.496
crash	1.000	-7.189	-1.087
glove	1.000	-9.370	-1.333
bean	1.000	-7.972	-1.632
opportunity	1.000	-3.405	-0.933
score	1.000	-3.613	-1.277
touch	1.000	-4.344	-0.521
dragon	1.000	-8.523	-1.596
everything	1.000	-2.934	-0.256
professor	1.000	-5.998	-1.118
useful	1.000	-4.052	-1.142
give	1.000	-2.307	-0.253
nervous	1.000	-5.432	-0.761
mercy	1.000	-9.013	-1.255
female	1.000	-5.092	-0.887
bakery	1.000	-7.850	-1.683
smell	1.000	-5.862	-0.547
life	1.000	-2.008	-0.166
bruise	1.000	-10.335	-1.647
gear	1.000	-8.751	-1.108
swallow	1.000	-8.831	-1.106

want	566.735	-1.770	-0.163
news	566.312	-2.751	-0.585
cheap	565.835	-5.404	-0.960
boat	565.338	-4.587	-0.982
join	564.655	-4.680	-0.773
protein	564.475	-9.156	-2.013
kid	562.831	-5.464	-0.669
same	562.102	-2.747	-0.211
agree	561.185	-3.694	-0.937
yellow	554.798	-6.279	-0.572
baby	553.380	-4.600	-0.565
court	551.058	-5.136	-0.933
final	540.662	-4.806	-0.709

rough	1.000	-5.031	-0.854
responsibility	1.000	-4.927	-1.112
beauty	1.000	-5.627	-0.770
selfish	1.000	-7.044	-1.500
obey	1.000	-7.007	-1.546
strategy	1.000	-5.578	-1.508
couple	1.000	-4.133	-0.475
rush	1.000	-5.830	-0.822
cry	1.000	-6.850	-0.655
freedom	1.000	-5.257	-1.042
zip	1.000	-9.849	-1.712
counter	1.000	-7.321	-0.820
opening	1.000	-5.207	-0.687

Appendix G: Correlations between Semantic Context Indices and other Related Indices – LSA (top) and Word2vec (bottom)

	EF-CAMDAT - Frequency Log	*EF-CAMDATA Range Log*	*MRC Concre-teness*	*MRC Familia-rity*	*MRC Imagea-bility*	*MRC Meaning-fulness*	*Context-tual Distincti-veness*[b]	*Semantic Diversity*[c]	*Age of Acquisition*[d]	*LSA*[e] *Average All Cosine*	*LSA Average Top Three Cosine*	*LSA Max Similarity Cosine*
Number of overlapping words	16031	16031	3865	4384	4313	2505	7092	12872	13615	4995	4978	4978
EF-CAMDAT LSA - Highest cosine similarity	0.338***	0.324***	0.142***	0.138***	0.146***	0.109***	-0.073***	-0.034***	-0.129***	0.113***	0.102***	0.089***
EF-CAMDAT LSA - Second highest cosine	0.351***	0.341***	0.136***	0.136***	0.141***	0.112***	-0.070***	-0.020*	-0.135***	0.108***	0.098***	0.082***
EF-CAMDAT LSA - Third highest cosine	0.354***	0.346***	0.130***	0.136***	0.135***	0.110***	-0.070***	-0.012	-0.136***	0.101***	0.091***	0.076***
EF-CAMDAT LSA - Average top three cosine	0.351***	0.340***	0.137***	0.138***	0.142***	0.111***	-0.072***	-0.022*	-0.135***	0.108***	0.097***	0.083***
EF-CAMDAT LSA - Average of cosine	0.722***	0.716***	0.063***	0.417***	0.088***	0.262***	-0.298***	0.136***	-0.340***	0.013	0.033*	0.030*
EF-CAMDAT LSA - Slope	0.305***	0.304***	-0.014	0.042*	-0.017	-0.01	0.093***	0.105***	-0.058***	-0.014	-0.002	-0.015
EF-CAMTAT LSA – Number of cosines above .3	0.110***	0.105***	0.168***	-0.004	0.195***	0.058*	0.026*	-0.093***	-0.101***	0.115***	0.128***	0.109***
EF-CAMDAT LSA - Average cosine above .3	0.318***	0.313***	0.074***	0.128***	0.072***	0.090***	-0.068***	0.015***	-0.114***	0.062***	0.052***	0.036*

[a] From the MRC database (Coltheart, 1981); [b] From McDonald and Shillcock (2001); [c] From Hoffman et al. (2013); [d]Kuperman et al. (2012); [e] From Landauer et al. (1998); *** $p < .0005$, ** $p < .005$, * $p < 0.05$, $p > 0.05$

	EF-CAMDAT - Frequency Log	*EF-CAMDATA Range Log*	*MRC Concre-teness*	*MRC Familia-rity*	*MRC Imagea-bility*	*MRC Meaning-fulness*	*Contex-tual Distincti-veness*[b]	*Semantic Diversity*[c]	*Age of Acquisition*[d]	*LSA*[e] *Average All Cosine*	*LSA Average Top Three Cosine*	*LSA Max Similarity Cosine*
Number of overlapping words	16031	16031	3865	4384	4313	2505	7092	12872	13615	4995	4978	4978
EF-CAMDAT Word2vec - Highest cosine similarity	0.093***	0.083**	0.327***	-0.011	0.324***	0.324***	0.013	-0.232***	-0.121***	0.230***	0.209***	0.188***
EF-CAMDAT Word2vec - Second highest cosine	-0.013	-0.023	0.359***	-0.066***	0.349***	0.349***	0.052***	-0.271***	-0.096***	0.243***	0.221***	0.200***
EF-CAMDAT Word2vec -Third highest cosine	-0.077***	-0.087***	0.369***	-0.097***	0.356***	0.356***	0.073***	-0.289***	-0.078***	0.242***	0.226***	0.201***
EF-CAMDAT Word2vec - Average top three cosine	0.003	-0.007	0.360***	-0.059***	0.352***	0.352***	0.046***	-0.269***	-0.101***	0.244***	0.225***	0.202***
EF-CAMDAT Word2vec - Average of cosine	0.775***	0.773***	0.057***	0.436***	0.089***	0.089***	-0.266***	0.166***	-0.353***	0.026	0.040**	0.036**
EF-CAMDAT Word2vec -Slope	-0.295***	-0.294***	0.002	-0.275***	-0.022	-0.022	0.281***	-0.098***	0.195***	-0.004	-0.004	-0.016
EF-CAMTAT Word2vec – Number of cosines above .3	-0.493***	-0.497***	0.158***	-0.302***	0.137***	0.137***	0.111***	-0.208***	0.106***	0.071***	0.063***	0.059***
EF-CAMDAT Word2vec - Average cosine above .3	-0.055***	-0.063***	0.255***	-0.014	0.230***	0.230***	-0.052***	-0.185***	-0.069***	0.167***	0.154***	0.141***

[a] From the MRC database (Coltheart, 1981); [b] From McDonald and Shillcock (2001); [c] From Hoffman et al. (2013); [d]Kuperman et al. (2012); [e] From Landauer et al. (1998); *** p < .0005, ** p < .005, * p < 0.05, p > 0.05

Appendix H: EF-CAMDAT and TASA Model Comparison Statistics

EF-CAMDAT Independent Model Comparisons

	Model description			Test against prior model	
Model	***Fixed Effects***	***Random Effects***	***AIC***	***Statistic***	***p***
1	None	language	1219.7		
2	Model 1 + EF-CAMDAT W2V - Average of all cosines	language	1128.5	$X^2(1) = 93.262$	<.005
3	Model 2 + EF-CAMDAT LSA - Number of cosines above .3	language	1123.2	$X^2(1) = 7.224$	0.01
4	Model 3 + EF-CAMDAT LSA – Average of cosines above .3	language	1121.5	$X^2(1) = 3.755$	0.06
5	Model 3 + EF-CAMDAT W2V - Number of cosines above .3	language	1125.2	$X^2(1) = 0.009$	0.92
6	Model 3 + EF-CAMDAT W2V - Average cosine above .3	language	1124.2	$X^2(1) = 0.991$	0.31

EF-CAMDAT Integrated Model Comparisons

	Model description			Test against prior model	
Model	***Fixed Effects***	***Random Effects***	***AIC***	***Statistic***	***p***
1	None	language	1556.2		
2	Model 1 + EF-CAMDAT W2V - Slope	language	1513.5	$X^2(1) = 44.661$	<.005
3	Model 2 + EF-CAMDAT LSA - Slope	language	1515.5	$X^2(1) = 0.0696$	0.95
4	Model 2 + EF-CAMDAT W2V – Number of cosines above .3	language	1511.5	$X^2(1) = 3.946$	0.05
5	Model 4 + EF-CAMDAT LSA – Average of all cosines[a]	language	1509.3	$X^2(1) = 4.189$	0.04
6	Model 2 + EF-CAMDAT Word2vec - Highest cosine similarity	language	1508.0	$X^2(1) = 1.300$	<.005

[a] Model suffered from suppression

TASA Independent Model Comparisons

Model description				Test against prior model	
Model	***Fixed Effects***	***Random Effects***	***AIC***	***Statistic***	***p***
1	None	language	1219.7		
2	Model 1 + TASA LSA - Average all cosine	language	1218.2	$X^2(1) = 3.546$	0.06
3	Model 1 + TASA LSA - Max similarity cosine	language	1218.2	$X^2(1) = 3.504$	0.06

TASA Integrated Model Comparisons

Model description				Test against prior model	
Model	***Fixed Effects***	***Random Effects***	***AIC***	***Statistic***	***p***
1	None	language	1556.2		
2	Model 1 + TASA LSA - Average top three cosine	language	1549.8	$X^2(1) = 8.3365$	<.005
3	Model 2 + TASA LSA - Average all cosine	language	1536.4	$X^2(1) = 15.455$	<.005

Combined Independent Model Comparisons

Model description				Test against prior model	
Model	***Fixed Effects***	***Random Effects***	***AIC***	***Statistic***	***p***
1	None	language	1219.7		
2	Model 1 + EF-CAMDAT W2V - Average of all cosines	language	1128.5	$X^2(1) = 93.262$	<.005
3	Model 2 + EF-CAMDAT LSA - Number of cosines above .3	language	1123.2	$X^2(1) = 7.224$	0.01
4	Model 3 + EF-CAMDAT LSA – Average of cosines above .3	language	1121.5	$X^2(1) = 3.755$	0.06
5	Model 3 + EF-CAMDAT W2V - Number of cosines above .3	language	1125.2	$X^2(1) = 0.009$	0.92
6	Model 3 + EF-CAMDAT W2V - Average cosine above .3	language	1124.2	$X^2(1) = 0.991$	0.31
7	Model 3 + TASA LSA – Average of all cosines	language	1124.7	$X^2(1) = 0.561$	0.45
8	Model 3 + TASA LSA – Maximum similarity cosine	language	1122.4	$X^2(1) = 2.832$	0.09

Combined Integrated Model Comparisons

Model description				Test against prior model	
Model	***Fixed Effects***	***Random Effects***	***AIC***	***Statistic***	***p***
1	None	language	1556.2		
2	Model 1 + EF-CAMDAT W2V - Slope	language	1513.5	$X^2(1) = 44.661$	<.005
3	Model 2 + EF-CAMDAT LSA - Slope	language	1515.5	$X^2(1) = 0.0696$	0.95
4	Model 2 + EF-CAMDAT W2V – Number of cosines above .3	language	1511.5	$X^2(1) = 3.946$	0.05
5	Model 4 + EF-CAMDAT LSA – Average of all cosines[a]	language	1509.3	$X^2(1) = 4.189$	0.04
6	Model 4 + TASA LSA – Average top three cosine	language	1506.0	$X^2(1) = 7.588$	0.01
7	Model 6 + EF-CAMDAT W2V – Highest cosine similarity	language	1505.8	$X^2(1) = 2.190$	0.13

[a] Model suffered from suppression effect

Appendix I: Individual Output with Semantic Context Indices – Independent Task

Independent essay – Score: 4.5 – Topic: career choice

The importance of like or dislike of certain subjets is commonly known. If there is a choice students should always decide for the more favourable subjects and never take some classes they do not agree with for whatever reasons.

For example it does not make to much sense for sports affected students which plan a career in baseball to take a course teaching corporate finance or statistics. As well it is not really helpful for a becoming history teacher to learn more about supply chain management. This does not support the idea of diversity in employment nad is not effectiv on the students side of learning and moving forward.

But sometimes the ways of becoming a professional in certain areas a student wants to work in later are bumpy. Students therefore have to take classes which do not agree with their understanding of the future profession on the first sight. But these subjects pay of later in their careers. To state an example it is absolutly nessacary for future engineers to gain some additional knowledge in business related areas even if they would prefer to do some more calculations on enginees or structure. Especially in todays industries this is a feature of high value and should be a basic component of a resume.

In conclusion I can smmerize that its always the smarter choice to go for subjects a student likes the most but in nowadays lifes this is a luxery not everybody can afford. Sometimes it is essential to study things which are not the most favourite ones to achieve to the desired target.

Independent essay – Score: 1.5 – Topic: career choice

in my opinion, choose subjects to prepare for a job is better than the interested. when the people became adults, they should earn the money from the jobs, not interested. choose subjects to prepare for the job because the subjects means future, the it is for the nice life and it is more saveful for the future.

first, we should think about the future. a good job is not easy to find. now we learned that we should use these knowage to deal with our job. so we choose the subjects stand for the future.

every body wants a nice life, relax and comfortable,not every day to do the hard work. if as young, people choose a good subject that propobly is good to find a good job. for example you have a famly factory, then you can continue the family bussiness. not do you interested to earn the money. that probely is hard.

choose a good subject that would be a long time. you should be careful. when people think about it deeply and decide the course, it would be more savefulthan the interested. because you thought and asked that is a plan. it is good for future. if you just like your interested, you did not think and aked, just od it, that probely is dangours.do everything we need a plan that is more saeful.

choose the subject to prepare for a job is better. save , have anice life and

Semantic Index Scores calculated for the High-scored and Low-scored Essays

Scores	***EF-CAMDAT W2V – Average of all cosines***	***EF-CAMDAT LSA - Number of cosines above .3 threshold***	***TASA – Average of all cosines (NOT SIGNIFICANT)***
4.5	0.871	121.737	0.160
1.5	0.904	100.900	0.149

High-scored essay (4.5)

Lemmas (type)	*Token count*	*EF-CAMDAT W2V - Average of all cosines*	*Lemmas (type)*	*Token count*	*EF-CAMDAT LSA - Number of cosines above .3*	*Lemmas (type)*	*Token count*	*TASA LSA - Average of all cosines (NOT SIG.)*
bumpy	1	0.268	diversity	1	536	baseball	1	0.553
doe	2	0.611	corporate	1	393	employment	1	0.437
todays	1	0.658	state	1	369	sports	1	0.427
commonly	1	0.658	industry	1	343	high	1	0.393
dislike	1	0.764	supply	1	294	student	2	0.381
certain	2	0.768	sport	1	289	students	4	0.381
sight	1	0.776	idea	1	280	forward	1	0.370
feature	1	0.790	choice	2	275	supply	1	0.369
desire	1	0.800	professional	1	253	teacher	1	0.360
conclusion	1	0.803	management	1	221	career	1	0.346
corporate	1	0.807	chain	1	216	careers	1	0.346
sense	1	0.808	engineer	1	212	later	2	0.327
supply	1	0.811	target	1	205	choice	2	0.281
diversity	1	0.824	teacher	1	205	agree	2	0.270
structure	1	0.829	bumpy	1	201	are	2	0.263
statistic	1	0.833	component	1	187	be	1	0.263
importance	1	0.833	conclusion	1	172	is	8	0.263
chain	1	0.836	sense	1	170	achieve	1	0.251
calculation	1	0.845	class	2	168	gain	1	0.247
target	1	0.845	achieve	1	164	decide	1	0.245
afford	1	0.847	career	2	158	sense	1	0.235
understanding	1	0.847	area	2	153	smarter	1	0.232
everybody	1	0.853	statistic	1	153	industries	1	0.230
basic	1	0.853	desire	1	150	value	1	0.205
reason	1	0.856	importance	1	148	well	1	0.192
component	1	0.858	feature	1	143	classes	2	0.188
state	1	0.859	knowledge	1	141	management	1	0.187
helpful	1	0.862	support	1	141	dislike	1	0.184
additional	1	0.864	doe	2	137	pay	1	0.183
employment	1	0.865	profession	1	137	profession	1	0.179
especially	1	0.865	certain	2	136	affected	1	0.177
finance	1	0.865	teach	1	136	reasons	1	0.173
affect	1	0.867	example	2	132	professional	1	0.159
baseball	1	0.870	employment	1	127	basic	1	0.156
essential	1	0.873	basic	1	121	becoming	2	0.154
choice	2	0.877	understanding	1	120	bumpy	1	0.151
nowadays	1	0.878	history	1	119	like	1	0.137
value	1	0.878	structure	1	115	likes	1	0.137
gain	1	0.879	finance	1	114	sight	1	0.133

Low-scored essay (1.5)

Lemmas	*Token count*	*Average of EF-CAMDAT W2V - Average of all cosines*	*Lemmas(types)*	*Token Count*	*EF-CAMDAT LSA - Number of cosines above .3*	*Lemmas (types)*	*Token count*	*TASA LSA - average of all cosines (NOT SIG.)*
od	1	0.495	choose	6	318	earn	2	0.348
mean	1	0.807	stand	1	249	be	3	0.263
deeply	1	0.821	deal	1	227	is	11	0.263
body	1	0.840	factory	1	213	factory	1	0.252
interest	5	0.850	opinion	1	212	decide	1	0.245
prepare	3	0.853	deeply	1	207	think	3	0.240
stand	1	0.855	nice	2	205	hard	2	0.230
deal	1	0.876	young	1	202	choose	6	0.223
continue	1	0.879	body	1	168	day	1	0.222
example	1	0.886	relax	1	161	adults	1	0.196
adult	1	0.888	comfortable	1	157	money	2	0.184
better	2	0.888	interest	5	156	good	6	0.181
subject	7	0.888	easy	1	146	life	3	0.175
young	1	0.889	example	1	132	stand	1	0.172
careful	1	0.889	mean	1	128	job	6	0.170
earn	2	0.891	save	1	104	jobs	1	0.170
factory	1	0.898	decide	1	98	asked	1	0.168
choose	6	0.900	long	1	86	comfortable	1	0.159
relax	1	0.900	earn	2	81	became	1	0.154
comfortable	1	0.902	hard	2	79	opinion	1	0.153
decide	1	0.902	prepare	3	63	relax	1	0.144
day	1	0.906	subject	7	60	body	1	0.141
need	1	0.908	careful	1	58	like	1	0.137
nice	2	0.910	course	1	51	easy	1	0.132
save	1	0.911	family	1	46	plan	2	0.131
like	1	0.911	adult	1	44	save	1	0.128
ask	1	0.914	ask	1	40	subject	3	0.122
opinion	1	0.915	future	5	34	subjects	4	0.122
want	1	0.918	money	2	32	can	1	0.118
easy	1	0.920	plan	2	32	find	2	0.114
time	1	0.920	learn	1	25	time	1	0.112
use	1	0.923	continue	1	21	careful	1	0.109
long	1	0.925	better	2	19	people	3	0.106
plan	2	0.927	life	3	13	learned	1	0.102
family	1	0.928	find	2	8	family	1	0.100
find	2	0.928	day	1	6	nice	2	0.099
good	6	0.930	like	1	6	prepare	3	0.082
course	1	0.931	need	1	6	just	2	0.082
think	4	0.932	want	1	4	work	1	0.081

idea	1	0.885	reason	1	112	plan	1	0.131
example	2	0.886	calculation	1	101	business	1	0.128
history	1	0.886	smart	1	101	history	1	0.128
support	1	0.887	decide	1	98	never	1	0.124
subject	3	0.888	afford	1	96	prefer	1	0.124
agree	2	0.890	prefer	1	96	subjects	3	0.122
sport	1	0.895	baseball	1	93	engineers	1	0.122
smart	1	0.896	high	1	87	can	2	0.118
relate	1	0.896	late	2	86	known	1	0.112
profession	1	0.900	student	6	85	knowledge	1	0.110
decide	1	0.902	helpful	1	82	feature	1	0.109
professional	1	0.908	agree	2	81	sometimes	2	0.103
like	2	0.911	affect	1	80	learn	1	0.102
prefer	1	0.912	essential	1	75	learning	1	0.102
achieve	1	0.913	additional	1	71	chain	1	0.088
know	1	0.916	sight	1	70	target	1	0.085
want	1	0.918	relate	1	69	idea	1	0.084
late	2	0.919	gain	1	63	work	1	0.081
industry	1	0.919	commonly	1	62	example	2	0.078
class	2	0.920	nowadays	1	60	future	2	0.078
teach	1	0.923	subject	3	60	certain	2	0.077
thing	1	0.924	dislike	1	57	always	2	0.075
resume	1	0.925	course	1	51	course	1	0.075
engineer	1	0.927	resume	1	47	even	1	0.073
student	6	0.927	business	1	45	structure	1	0.071
plan	1	0.927	especially	1	43	component	1	0.069
area	2	0.930	everybody	1	42	helpful	1	0.069
course	1	0.931	value	1	39	state	1	0.069
teacher	1	0.932	future	2	34	moving	1	0.062
knowledge	1	0.933	plan	1	32	areas	2	0.056
business	1	0.934	todays	1	32	statistics	1	0.055
way	1	0.934	pay	1	30	study	1	0.055
management	1	0.936	learn	2	25	side	1	0.050
future	2	0.937	way	1	25	things	1	0.048
work	1	0.941	forward	1	22	support	1	0.045
career	2	0.942	study	1	17	do	3	0.040
high	1	0.942	like	2	6	does	2	0.040
pay	1	0.944	want	1	4	ways	1	0.037
forward	1	0.945	know	1	3	with	2	0.032
learn	2	0.948	thing	1	1	have	1	0.032
study	1	0.953				conclusion	1	0.030
						wants	1	0.027
						go	1	0.023
						desired	1	0.022
						take	3	0.022
						much	1	0.020
						make	1	0.014
						ones	1	0.014

hard	2	0.935	time	1	2	interested	5	0.081
future	5	0.937	use	1	2	example	1	0.078
people	3	0.939	od	1	1	future	5	0.078
job	7	0.940				course	1	0.075
work	1	0.941				means	1	0.066
life	3	0.944				deal	1	0.062
learn	1	0.948				need	1	0.058
money	2	0.948				use	1	0.052
						now	1	0.052
						did	1	0.040
						do	3	0.040
						with	1	0.032
						have	2	0.032
						wants	1	0.027
						long	1	0.014
						everything	1	0.005

Appendix J: **Individual Output with Semantic Context Indices – Integrated Task**

Integrated essay – Score: 4.5 – Topic: bird migration

The lecture tries to disprove that each of the three theories given in the reading passage about bird's navigation abilities can be a complete explanation for bird's navigation abilities.

First, the theory is discussed, that birds can navigate just with the help of celestial objects like the sun or stars. The lecturer mentioned that even if celestial objects are not visible (for example, if they're hidden by clouds), birds can still find their way. This fact shows that the celestial objects theory can't be an all-explaining theory.

Second, the lecturer disporves the fact that a bird's navigation only relies on remembering landmarks. He mentioned experiments that were made, in which birds were able to find their way back home through completely unknown teritorry.

To disprove that the third fact alone can be a explanation for bird's navigation abilities, the lecturer speaks about the fact, that it's impossible to find a certain place just with the help of a compas-like device. So even if birds have crystals of the mineral magnetite embedded in their beaks, this feature alone can't guide them to specific location.

But in the endaccording to the lecturer, all three theories combined could be a reasonable explanation for the fact, that birds find their way home.

Integrated essay – Score: 1.5 – Topic: bird migration

Birds are very accurate at navigating long distances. There is three principal theories about how the ability of birds of traveling long distances is so accurate. But all of these are not fully true, because each of them has limations.

The firts theory says that the bird travels in reference to the Sun by the day, and to the stars by night. Since both, the Sun and the stars, are not visible at all times, the birds sometimes get lost and start following the wrong star and end up going the wrong way.

The second theory claims that birds navigate by landmarks such as mountains, coastlines or rivers. Birds memorize this places, so that is how they get orientated. There is a region on the birds called hippocampal region, and when it gets damaged, the bird cannot use its ability as well, so its memory and ability to navigate gets worse too.

The third and last theory holds that birds use like an internal compass that responds to Earth's magnetic field. But a bird is not like a human that knows where he is, birds have self-orientation that sometimes fails on them, so that is why they lose track of where they are, soy they get lost and can not use the compass anymore.

In conclusion, birds could be really good and accurate at flying long distances, but sometimes the stars, their memory or internal compass coudl play tricks on them and make them get lost.

Semantic Index Scores calculated for the High-scored and Low-scored Essays

Score	EF-CAMDAT W2V - Slope	EF-CAMDAT W2V - Number of cosines above .3	EF-CAMDAT W2V - Highest cosine word similarity	TASA LSA - Average all cosine	TASA LSA - Average top three cosine
4.5	0.029	353.541	0.565	0.151	0.171
1.5	0.026	332.688	0.595	0.203	0.194

High-scored essay (4.5)													
Lemmas (type)	***Token count***	***EF-CAMDAT W2V - Slope***	***Lemmas (type)***	***Token count***	***EF-CAMDAT W2V- Number of cosines above .3***	***Lemmas (type)***	***Token count***	***EF-CAMDAT W2V - Highest cosine similarity***	***Lemmas (type)***	***Token count***	***TASA LSA - Average of all cosines***	***Lemmas (type)***	***Token count***	***TASA LSA - Average top three cosine***
visible	1	0.104	beak	1	3566	mention	2	0.334	birds	5	0.394	have	1	0.01
beak	1	0.099	celestial	3	2579	certain	1	0.387	remembering	1	0.318	crystals	1	0.03
crystal	1	0.078	crystal	1	1534	guide	1	0.398	experiments	1	0.299	are	1	0.04
landmark	1	0.074	bird	5	1361	fact	5	0.407	abilities	3	0.265	be	4	0.04
passage	1	0.067	mineral	1	1042	impossible	1	0.412	are	1	0.263	is	1	0.04
object	3	0.056	cloud	1	949	second	1	0.428	be	4	0.263	were	2	0.04
experiment	1	0.052	landmark	1	833	rely	1	0.449	is	1	0.263	can	4	0.05
explanation	3	0.048	embed	1	392	passage	1	0.455	were	2	0.263	made	1	0.05
feature	1	0.046	visible	1	389	place	1	0.457	theories	2	0.252	objects	3	0.05
certain	1	0.044	navigation	4	255	unknown	1	0.461	theory	3	0.252	guide	1	0.06
rely	1	0.043	device	1	254	try	1	0.467	mineral	1	0.240	with	2	0.07
combine	1	0.041	lecturer	4	239	sun	1	0.469	specific	1	0.192	place	1	0.08
unknown	1	0.041	ability	3	231	able	1	0.483	location	1	0.180	relies	1	0.08
navigation	4	0.039	hide	1	174	find	4	0.497	lecture	1	0.172	way	3	0.08
navigate	1	0.037	theory	5	163	visible	1	0.497	still	1	0.170	discussed	1	0.1
mineral	1	0.036	explanation	3	158	help	2	0.501	alone	2	0.169	feature	1	0.1
completely	1	0.035	specific	1	155	specific	1	0.504	complete	1	0.150	just	2	0.1
reasonable	1	0.033	navigate	1	149	way	3	0.509	fact	5	0.138	able	1	0.11
star	1	0.032	combine	1	148	star	1	0.514	like	1	0.137	explanation	3	0.11
fact	5	0.031	lecture	1	139	explanation	3	0.518	relies	1	0.129	help	2	0.11
mention	2	0.027	star	1	139	combine	1	0.538	discussed	1	0.126	shows	1	0.11
discuss	1	0.026	feature	1	103	reasonable	1	0.539	unknown	1	0.118	certain	1	0.12
celestial	3	0.025	object	3	89	crystal	1	0.539	can	4	0.118	combined	1	0.13
theory	5	0.025	unknown	1	87	discuss	1	0.549	device	1	0.117	even	2	0.13
embed	1	0.024	complete	1	65	complete	1	0.551	disprove	2	0.116	passage	1	0.13
impossible	1	0.023	rely	1	63	embed	1	0.552	speaks	1	0.115	unknown	1	0.14
sun	1	0.023	way	3	62	experiment	1	0.555	impossible	1	0.115	device	1	0.15
second	1	0.021	location	1	61	remember	1	0.558	find	4	0.114	example	1	0.17
specific	1	0.021	read	1	61	object	3	0.560	hidden	1	0.113	home	2	0.17
remember	1	0.021	reasonable	1	61	cloud	1	0.571	feature	1	0.109	second	1	0.17
lecture	1	0.020	sun	1	59	read	1	0.575	passage	1	0.107	back	1	0.18
able	1	0.020	experiment	1	56	celestial	3	0.585	able	1	0.097	lecture	1	0.2
guide	1	0.020	discuss	1	53	location	1	0.586	tries	1	0.095	complete	1	0.21
complete	1	0.019	help	2	53	example	1	0.589	guide	1	0.095	disprove	2	0.21
try	1	0.019	speak	1	51	home	2	0.594	explanation	3	0.094	impossible	1	0.21
lecturer	4	0.019	able	1	37	feature	1	0.595	second	1	0.091	like	1	0.21
bird	5	0.018	place	1	37	hide	1	0.599	just	2	0.082	mineral	1	0.21
example	1	0.017	passage	1	35	navigation	4	0.609	home	2	0.079	specific	1	0.21

cloud	1	0.016	like	1	31	beak	1	0.624	example	1	0.078	theories	2	0.21
like	1	0.016	remember	1	22	completely	1	0.628	certain	1	0.077	theory	3	0.21
home	2	0.015	try	1	21	speak	1	0.639	back	1	0.077	experiments	1	0.23
ability	3	0.013	certain	1	20	theory	5	0.660	even	2	0.073	location	1	0.23
device	1	0.013	guide	1	19	lecture	1	0.665	combined	1	0.070	tries	1	0.23
hide	1	0.012	fact	5	15	navigate	1	0.674	shows	1	0.063	alone	2	0.24
way	3	0.012	example	1	13	bird	5	0.686	objects	3	0.059	find	4	0.24
read	1	0.012	home	2	13	device	1	0.688	help	2	0.055	hidden	1	0.24
help	2	0.011	impossible	1	13	lecturer	4	0.693	place	1	0.042	speaks	1	0.25
place	1	0.011	completely	1	9	like	1	0.708	way	3	0.037	still	1	0.25
speak	1	0.010	find	4	8	ability	3	0.732	with	2	0.032	fact	5	0.28
find	4	0.010	second	1	7	landmark	1	0.740	have	1	0.032	abilities	3	0.33
location	1	0.007	mention	2	6	mineral	1	0.855	crystals	1	0.024	remembering	1	0.43
									made	1	0.014	birds	5	0.47

Low-scored essay (1.5)

Lemmas (type)	*Token count*	*EF-CAMDAT W2V - Slope*	*Lemmas (type)*	*Token count*	*EF-CAMDAT W2V-Number of cosines above .3*	*Lemmas (type)*	*Token count*	*EF-CAMDAT W2V - Highest cosine similarity*	*Lemmas (type)*	*Token count*	*TASA LSA - Average of all cosines*	*Lemmas (type)*	*Token count*	*TASA LSA - Average top three cosine*
visible	1	0.104	magnetic	1	2090	hold	1	0.381	compass	3	0.457	how	2	0
trick	1	0.076	soy	1	1775	follow	1	0.415	birds	8	0.394	why	1	0
landmark	1	0.074	bird	12	1361	second	1	0.428	bird	3	0.394	has	1	0.01
soy	1	0.073	landmark	1	833	conclusion	1	0.439	flying	1	0.347	have	1	0.01
principal	1	0.063	river	1	775	fail	1	0.448	human	1	0.341	there	2	0.01
claim	1	0.062	coastline	1	735	end	1	0.456	TRUE	1	0.313	are	4	0.04
fully	1	0.055	mountain	1	673	distance	3	0.456	worse	1	0.295	be	1	0.04
conclusion	1	0.047	compass	3	400	place	1	0.457	mountains	1	0.281	is	7	0.04
compass	3	0.046	visible	1	389	fully	1	0.464	lose	1	0.268	can	1	0.05
memorize	1	0.042	internal	2	347	sun	2	0.469	lost	3	0.268	going	1	0.05
accurate	3	0.040	accurate	3	241	reference	1	0.471	ability	3	0.265	make	1	0.05
track	1	0.038	ability	3	231	time	1	0.480	are	4	0.263	memory	2	0.07
navigate	3	0.037	region	2	226	compass	3	0.481	be	1	0.263	long	3	0.08
orientate	1	0.037	damage	1	221	long	3	0.489	is	7	0.263	places	1	0.08
reference	1	0.034	memorize	1	177	principal	1	0.490	track	1	0.260	way	1	0.08
magnetic	1	0.033	orientate	1	170	traveling	1	0.493	theories	1	0.252	conclusion	1	0.09
star	4	0.032	theory	4	163	visible	1	0.497	theory	3	0.252	says	1	0.09
memory	2	0.029	navigate	3	149	wrong	2	0.502	region	2	0.244	fails	1	0.1
follow	1	0.029	star	4	139	way	1	0.509	wrong	2	0.233	travels	1	0.1
damage	1	0.028	fly	1	133	star	4	0.514	rivers	1	0.225	claims	1	0.12
hold	1	0.027	play	1	116	human	1	0.519	day	1	0.222	use	3	0.12
internal	2	0.027	reference	1	106	claim	1	0.520	distances	3	0.221	well	1	0.12
anymore	1	0.025	wrong	2	101	magnetic	1	0.521	night	1	0.206	damaged	1	0.13
lose	4	0.025	bad	1	100	know	1	0.521	play	1	0.205	day	1	0.13
coastline	1	0.025	field	1	96	track	1	0.526	well	1	0.192	responds	1	0.14

theory	4	0.025	human	1	81	day	1	0.542	accurate	3	0.182	last	1	0.15
respond	1	0.023	claim	1	78	memory	2	0.551	good	1	0.181	worse	1	0.15
sun	2	0.023	use	3	76	orientate	1	0.552	principal	1	0.177	sometimes	3	0.16
second	1	0.021	conclusion	1	73	accurate	3	0.557	damaged	1	0.156	region	2	0.17
end	1	0.020	traveling	1	64	anymore	1	0.561	called	1	0.153	second	1	0.17
region	2	0.020	way	1	62	bad	1	0.568	tricks	1	0.145	times	1	0.17
wrong	2	0.019	sun	2	59	damage	1	0.583	travels	1	0.144	wrong	2	0.17
traveling	1	0.019	good	1	55	travel	1	0.584	field	1	0.140	end	1	0.19
river	1	0.018	night	1	51	fly	1	0.586	like	2	0.137	called	1	0.2
bird	12	0.018	trick	1	46	use	3	0.597	memory	2	0.136	principal	1	0.2
day	1	0.018	respond	1	45	memorize	1	0.607	can	1	0.118	like	2	0.21
like	2	0.016	memory	2	43	good	1	0.615	knows	1	0.112	lose	1	0.21
ability	3	0.013	distance	3	41	field	1	0.618	times	1	0.112	lost	3	0.21
long	3	0.013	place	1	37	trick	1	0.618	fails	1	0.109	theories	1	0.21
start	1	0.013	day	1	36	night	1	0.651	responds	1	0.109	theory	3	0.21
use	3	0.012	travel	1	32	region	2	0.656	sometimes	3	0.103	distances	3	0.22
way	1	0.012	like	2	31	respond	1	0.657	holds	1	0.101	field	1	0.22
mountain	1	0.012	know	1	27	theory	4	0.660	claims	1	0.101	start	1	0.22
know	1	0.012	track	1	27	navigate	3	0.674	second	1	0.091	knows	1	0.24
play	1	0.011	lose	4	26	play	1	0.674	says	1	0.079	night	1	0.24
night	1	0.011	principal	1	22	bird	12	0.686	end	1	0.076	tricks	1	0.24
field	1	0.011	follow	1	21	internal	2	0.706	last	1	0.070	good	1	0.25
bad	1	0.011	hold	1	20	like	2	0.708	start	1	0.063	accurate	3	0.27
place	1	0.011	time	1	20	river	1	0.728	use	3	0.052	mountains	1	0.27
fly	1	0.010	end	1	15	ability	3	0.732	places	1	0.042	rivers	1	0.27
distance	3	0.010	start	1	15	landmark	1	0.740	way	1	0.037	true	1	0.27
fail	1	0.010	fail	1	12	start	1	0.744	has	1	0.032	holds	1	0.3
time	1	0.010	anymore	1	11	lose	4	0.752	have	1	0.032	ability	3	0.33
good	1	0.009	fully	1	8	coastline	1	0.762	conclusion	1	0.030	flying	1	0.34
travel	1	0.007	second	1	7	mountain	1	0.775	going	1	0.023	compass	3	0.36
human	1	0.005	long	3	4	soy	1	0.827	make	1	0.014	human	1	0.37
									long	3	0.014	play	1	0.39
									why	1	0.009	track	1	0.42
									how	2	0.005	birds	8	0.47
												bird	3	0.47

Appendix K: Semantic Context Scores for the 100 Words with Higher and Lower RT and Accuracy scores

word	*L2 RT mean*	*EF-CAMDAT W2V average of all cosines*	*EF-CAMDAT W2V Third highest cosine word similarity*	*EF-CAMDAT LSA highest cosine word similarity*	*TASA LSA average all cosine*
sermon	1120.656	0.454	0.414	0.66	0.471
stud	1052.056	0.914	0.493	0.382	0.066
chapel	1033.796	0.62	0.618	0.563	0.205
verse	1031.21	0.746	0.595	0.488	0.357
shrub	1028.813	0.343	0.621	0.593	0.438
linen	1023.743	0.74	0.736	0.875	0.208
sow	1023.286	0.733	0.384	0.319	0.222
philosopher	1020.77	0.635	0.488	0.647	0.38
mower	1020.563	0.518	0.637	0.584	0.354
attend	1013.967	0.89	0.46	0.585	0.128
paddy	1012.067	0.218	0.469	0.549	0.057
gown	1010.856	0.645	0.637	0.443	0.329
tornado	989.865	0.884	0.73	0.923	0.568
curve	986.574	0.829	0.667	0.85	0.094
wag	986.043	0.093	0.652	0.595	0.629
giggle	984.904	0.204	0.626	0.467	0.413
cauliflower	966.843	0.248	0.632	0.616	0.23
tramp	965.346	0.292	0.476	0.696	0.035
continent	963.629	0.857	0.543	0.886	0.202
cricket	962.829	0.617	0.507	0.556	0.219
loser	962.725	0.81	0.493	0.739	0.356
noun	962.533	0.528	0.562	0.676	0.793
spouse	961.907	0.752	0.507	0.688	0.292
campus	961.131	0.81	0.495	0.851	0.259
sick	960.743	0.881	0.482	0.478	0.311
instance	958.24	0.792	0.415	0.505	0.07
expression	957.535	0.806	0.543	0.667	0.188
knit	956.908	0.655	0.441	0.635	0.197
pin	956.3	0.929	0.63	0.558	0.197
circumstance	954.857	0.785	0.419	0.887	0.198
mayor	954.38	0.692	0.491	0.959	0.247
payment	950.756	0.872	0.512	0.775	0.275
drag	949.374	0.766	0.646	0.8	0.151
bureau	948.125	0.714	0.46	0.648	0.132
dam	947.949	0.723	0.583	0.866	0.31
erect	947.914	0.464	0.609	0.645	0.161
mend	941	0.561	0.472	0.319	0.166
manufacture	940.635	0.92	0.66	0.862	0.176
border	939.792	0.775	0.512	0.889	0.179
frightened	936.254	0.886	0.718	0.741	0.182
intake	935.744	0.458	0.517	0.571	0.096
frightening	932.616	0.88	0.693	0.805	0.182
sword	929.791	0.595	0.567	0.957	0.204
lieutenant	928.18	0.486	0.539	0.757	0.318
tart	928.068	0.721	0.878	0.955	0.201
blonde	928.02	0.814	0.671	0.598	0.254
con	926.895	0.668	0.622	0.998	0.053
wardrobe	923.451	0.834	0.716	0.838	0.078
hum	922.762	0.646	0.63	0.636	0.295
vase	922.667	0.811	0.726	0.937	0.216
riot	922.255	0.584	0.458	0.531	0.206
wander	921.684	0.769	0.549	0.47	0.131
different	921.386	0.93	0.45	0.568	0.07

word	*L2 acc mean*	*EF-CAMDAT W2V Third highest cosine word similarity*	*EF-CAMDAT W2V average of all cosines*	*EF-CAMDAT W2V number of cosines above .3*	*EF-CAMDAT LSA highest cosine word similarity*	*EF-CAMDAT LSA slope*	*TASA LSA average top three cosine*
paddy	0.25	0.47	0.22	709	0.55	0.03	0.03
mousse	0.32	0.78	0.66	2100	0.88	0.12	0.06
muck	0.32	0.65	0.37	4418	0.35	0.03	0.12
sow	0.39	0.38	0.73	150	0.32	0.06	0.27
mare	0.45	0.48	0.57	1074	0.47	0.05	0.49
treble	0.45	0.59	0.42	3474	0.39	0.02	0.43
bog	0.5	0.52	0.53	1419	0.46	0.02	0.11
apex	0.6	0.42	0.53	80	0.7	0.01	0.03
gin	0.6	0.57	0.43	3478	0.36	0.02	0.3
wary	0.61	0.53	0.75	329	0.79	0.09	0.12
con	0.63	0.62	0.67	32	1	0.14	0.13
owe	0.63	0.39	0.68	40	0.94	0.13	0.48
tart	0.63	0.88	0.72	1588	0.95	0.12	0.31
wit	0.63	0.44	0.58	1306	0.41	0.09	0.17
atom	0.64	0.53	0.63	1572	0.69	0.05	0.35
hum	0.64	0.63	0.65	2565	0.64	0.05	0.27
dwarf	0.65	0.58	0.27	4546	0.55	0.07	0.12
ale	0.65	0.69	0.8	2926	0.48	0.03	0.28
mason	0.65	0.45	0.62	890	0.55	0.09	0.01
twig	0.65	0.51	0.14	1746	0.44	0.05	0.37
dam	0.67	0.58	0.72	494	0.87	0.1	0.39
digger	0.67	0.68	0.22	5260	0.48	0.04	0.23
gown	0.67	0.64	0.65	2055	0.44	0.07	0.26
linen	0.67	0.74	0.74	1273	0.88	0.11	0.33
mend	0.67	0.47	0.56	1359	0.32	0.03	0.26
swan	0.67	0.59	0.7	2167	0.85	0.12	0.36
triumph	0.67	0.51	0.32	2710	0.61	0.13	0.25
elm	0.68	0.52	0.59	271	0.99	0.15	0.45
filthy	0.68	0.65	0.8	2222	0.77	0.03	0.18
rifle	0.68	0.5	0.31	931	0.48	0.07	0.47
saucer	0.68	0.54	0.12	2432	0.54	0.05	0.28
cod	0.7	0.72	0.71	1008	0.57	0.08	0.28
gallop	0.7	0.64	0.37	4617	0.49	0.02	0.54
diminish	0.71	0.53	0.68	251	0.34	0.1	0.1
mop	0.71	0.59	0.68	2210	0.57	0.06	0.23
rake	0.71	0.6	0.37	3694	0.36	0.02	0.17
reel	0.71	0.49	0.33	1441	0.47	0.02	0.1
dart	0.71	0.59	0.76	497	0.65	0.05	0.07
bulb	0.72	0.59	0.7	537	0.56	0.1	0.24
feast	0.72	0.6	0.71	912	0.82	0.12	0.3
monarchy	0.72	0.5	0.22	1466	0.38	0.07	0.29
mower	0.72	0.64	0.52	5285	0.58	0.03	0.29
starve	0.72	0.53	0.8	304	0.76	0.09	0.42
width	0.72	0.8	0.91	1190	0.97	0.07	0.53
foam	0.74	0.55	0.59	1669	0.49	0.07	0.19
hay	0.74	0.43	0.55	287	0.95	0.16	0.37
hip	0.74	0.5	0.75	172	0.79	0.08	0.42
ton	0.74	0.44	0.79	147	0.46	0.1	0.36
vow	0.74	0.49	0.43	1772	0.66	0.11	0.18
bureau	0.75	0.46	0.71	144	0.65	0.05	0.09
crook	0.75	0.5	0.1	465	0.61	0.06	0.25
fist	0.75	0.29	0.73	2	0.44	0.06	0.22
pigeon	0.75	0.43	0.37	606	0.54	0.04	0.22

excitement	921.071	0.671	0.449	0.609	0.097
width	920.615	0.906	0.796	0.968	0.5
frequent	919.832	0.772	0.468	0.976	0.017
dense	919.823	0.636	0.588	0.52	0.12
burglary	919.624	0.192	0.558	0.998	0.208
hip	917.208	0.751	0.496	0.795	0.314
rape	914.646	0.85	0.733	0.929	0.183
climb	912.997	0.879	0.627	0.816	0.22
nun	912.774	0.427	0.561	0.447	0.152
stem	912.616	0.784	0.581	0.749	0.404
squeeze	911.907	0.718	0.731	0.814	0.096
quilt	910.537	0.562	0.677	0.832	0.229
helper	910.3	0.735	0.416	0.79	0.063
cabbage	905.94	0.667	0.836	0.901	0.241
scrap	904.072	0.716	0.462	0.413	0.199
choir	903.513	0.76	0.687	0.662	0.389
vague	902.931	0.532	0.41	0.36	0.108
shiver	901.414	0.681	0.597	0.492	0.344
perceive	900.915	0.684	0.47	0.42	0.079
confusion	900.444	0.701	0.539	0.965	0.033
lemonade	899.888	0.691	0.768	0.861	0.168
competition	897.796	0.844	0.475	0.565	0.225
accurate	896.342	0.8	0.54	0.787	0.182
sew	896.104	0.802	0.631	0.719	0.218
queen	892.984	0.803	0.608	0.613	0.097
slippery	892.478	0.761	0.592	0.575	0.206
yacht	890.303	0.834	0.721	0.542	0.228
street	890.225	0.871	0.517	0.694	0.149
graph	889.961	0.655	0.609	0.838	0.149
timber	888.78	0.401	0.61	0.546	0.309
cooking	888.468	0.875	0.493	0.981	0.191
scheme	888.262	0.69	0.388	0.916	0.071
sense	887.479	0.808	0.412	0.643	0.235
identical	887.06	0.412	0.448	0.663	0.216
flip	886.67	0.677	0.526	0.395	0.098
snore	885.474	0.224	0.587	0.867	0.363
sergeant	885.349	0.591	0.57	0.641	0.226
hen	885.094	0.881	0.611	0.943	0.485
mare	884.809	0.573	0.477	0.471	0.695
fence	884.593	0.719	0.583	0.485	0.167
diminish	884.42	0.685	0.533	0.338	0.07
bold	883.981	0.386	0.598	0.993	0.106
ditch	882.724	0.594	0.599	0.412	0.174
vodka	881.592	0.767	0.767	0.67	0.464
slab	880.846	0.294	0.588	0.556	0.192
brook	880.819	0.634	0.609	0.99	0.003
late	879.501	0.919	0.461	0.74	0.327
ocean	610.72	0.875	0.651	0.525	0.277
criminal	610.718	0.771	0.64	0.921	0.394
bear	610.549	0.824	0.395	0.677	0.214
tell	610.475	0.919	0.408	0.556	0.148
raw	610.392	0.852	0.571	0.697	0.09
fox	610.369	0.712	0.595	0.603	0.194
moving	610.108	0.578	0.406	0.419	0.062
foot	609.874	0.868	0.624	0.597	0.245
cell	609.816	0.874	0.562	0.958	0.127
glue	609.775	0.623	0.549	0.389	0.107
snake	609.437	0.922	0.746	0.768	0.271
single	609.308	0.776	0.292	0.655	0.108
class	609.246	0.92	0.555	0.81	0.188
press	609.214	0.872	0.472	0.635	0.218
hard	608.951	0.935	0.504	0.665	0.23
language	608.884	0.935	0.511	0.793	0.426
dance	608.822	0.886	0.609	0.907	0.275
weapon	608.334	0.829	0.601	0.754	0.218
city	607.948	0.933	0.583	0.694	0.1
nation	607.568	0.871	0.539	0.951	0.122
book	606.925	0.928	0.501	0.745	0.314
free	606.895	0.903	0.354	0.611	0.109

rub	0.75	0.61	0.77	1158	0.63	0.07	0.17
shore	0.75	0.61	0.83	1055	0.68	0.05	0.41
yacht	0.75	0.72	0.83	1345	0.54	0.06	0.13
bias	0.76	0.47	0.54	226	0.7	0.09	0.24
paw	0.76	0.45	0.48	730	0.37	-0.02	0.27
iron	0.76	0.53	0.75	525	0.97	0.12	0.29
lieutenant	0.76	0.54	0.49	2194	0.76	0.12	0.39
dust	0.77	0.64	0.86	1571	0.43	0.07	0.26
bra	0.78	0.47	0.4	1599	0.35	0.04	0.17
crease	0.78	0.46	0.55	863	0.52	0.01	0.32
ditch	0.78	0.6	0.59	4638	0.41	0.01	0.2
hen	0.78	0.61	0.88	1388	0.94	0.05	0.58
hymn	0.78	0.44	0.06	170	0.58	0.04	0.34
meadow	0.78	0.7	0.6	3348	0.64	0.03	0.33
tack	0.78	0.4	0.57	338	0.35	0.05	0.12
hut	0.78	0.58	0.83	2270	0.55	0.04	0.1
camel	0.79	0.68	0.91	2415	0.8	0.04	0.47
chalk	0.79	0.55	0.79	1667	0.53	0.04	0.16
essence	0.79	0.5	0.6	356	0.81	0.13	0.07
hedge	0.79	0.45	0.58	114	0.39	0.09	0.1
lamb	0.79	0.88	0.83	1718	0.96	0.1	0.45
loft	0.79	0.61	0.57	1411	0.47	0.1	0.26
miner	0.79	0.43	0.41	592	0.57	0.1	0.35
pinch	0.79	0.5	0.37	1053	0.61	0.1	0.2
prefer	0.79	0.41	0.91	19	0.76	0	0.12
sew	0.79	0.63	0.8	1111	0.72	0.04	0.23
tin	0.79	0.54	0.58	1640	0.39	0.06	0.34
tread	0.79	0.49	0.58	2337	0.51	0.04	0.1
wander	0.79	0.55	0.77	1540	0.47	0.05	0.26
barn	0.8	0.59	0.45	4121	0.38	0.04	0.51
binder	0.8	0.87	0.35	6573	0.41	0.04	0.07
disturb	0.8	0.53	0.82	170	0.49	0.06	0.09
fireplace	0.8	0.66	0.6	1518	0.82	0.12	0.34
floor	0.8	0.58	0.86	629	0.62	0.1	0.48
lodge	0.8	0.49	0.88	58	0.92	0.06	0.12
mental	0.8	0.57	0.85	250	0.66	0.09	0.24
merit	0.8	0.45	0.75	96	0.59	0.07	0.06
myth	0.8	0.63	0.3	733	0.76	0.03	0.3
noun	0.8	0.56	0.53	154	0.68	0.08	0.8
patch	0.8	0.51	0.44	2614	0.37	0.07	0.07
pint	0.8	0.54	0.82	553	0.64	0	0.28
praise	0.8	0.5	0.79	145	0.61	0.09	0.12
shrub	0.8	0.62	0.34	4724	0.59	0.01	0.46
slack	0.8	0.55	0.53	2664	0.79	0.07	0.07
snore	0.8	0.59	0.22	2968	0.87	0.1	0.35
stud	0.8	0.49	0.91	82	0.38	0.05	0.01
tea	0.8	0.56	0.82	373	0.86	0.07	0.19
permission	1	0.44	0.82	35	0.92	0.1	0.16
beware	1	0.38	0.63	86	0.5	0.09	0.08
decoration	1	0.69	0.86	609	0.74	0.08	0.13
cocktail	1	0.7	0.82	1010	0.77	0.1	0.29
clear	1	0.5	0.82	73	0.69	0.11	0.11
half	1	0.4	0.86	29	0.49	0.09	0.12
steak	1	0.67	0.75	523	0.96	0.11	0.35
philosophy	1	0.63	0.86	208	0.44	0.06	0.14
assistant	1	0.58	0.91	200	0.55	0.05	0.11
boss	1	0.47	0.89	32	0.59	0.05	0.16
pass	1	0.36	0.9	16	0.97	0	0.03
daily	1	0.35	0.87	15	0.96	0.11	0.39
girl	1	0.56	0.91	140	0.7	0.03	0.18
tonight	1	0.5	0.82	246	0.66	0.09	0.18
chair	1	0.58	0.85	210	0.82	0.09	0.53
dawn	1	0.52	0.78	953	0.49	0.07	0.01
profession	1	0.55	0.9	58	0.72	0.07	0.2
newspaper	1	0.49	0.89	68	0.74	0.04	0.65
peace	1	0.42	0.88	22	0.85	0.06	0.27
depression	1	0.66	0.86	485	0.59	0.07	0.14
temperature	1	0.56	0.93	392	0.73	0.05	0.35
blood	1	0.5	0.81	319	0.87	0.12	0.18

floor	606.5	0.862	0.578	0.62	0.284
agent	606.171	0.861	0.526	0.99	0.09
stock	605.529	0.869	0.484	0.541	0.09
broken	604.632	0.822	0.532	0.478	0.089
village	604.397	0.895	0.573	0.883	0.182
pen	603.938	0.804	0.563	0.989	0.295
grass	603.915	0.89	0.658	0.598	0.223
command	603.257	0.845	0.389	0.405	0.193
play	603.088	0.931	0.527	0.856	0.205
cow	603.064	0.9	0.717	0.908	0.5
angry	603.027	0.858	0.525	0.609	0.191
positive	602.702	0.864	0.444	0.66	0.705
knock	602.557	0.932	0.77	0.643	0.447
calendar	602.429	0.812	0.477	0.625	0.27
wipe	601.795	0.732	0.596	0.969	0.091
opera	601.743	0.836	0.519	0.99	0.013
chicken	601.525	0.875	0.803	0.889	0.185
box	600.813	0.736	0.461	0.906	0.083
pause	600.174	0.738	0.43	0.663	0.095
sky	600.147	0.871	0.561	0.926	0.364
wine	600.062	0.834	0.687	0.732	0.361
plane	598.751	0.886	0.542	0.871	0.333
hospital	598.627	0.917	0.565	0.825	0.395
thinking	598.362	0.873	0.459	0.913	0.24
advice	598.027	0.846	0.444	0.825	0.12
physical	597.882	0.822	0.477	0.71	0.297
use	597.879	0.923	0.451	0.331	0.052
local	597.861	0.816	0.334	0.735	0.125
office	597.177	0.899	0.365	0.722	0.18
hood	597.031	0.558	0.488	0.391	0.158
root	596.647	0.744	0.431	0.587	0.334
camera	595.575	0.849	0.526	0.704	0.53
imagination	595.147	0.781	0.464	0.602	0.184
beer	595.115	0.888	0.68	0.783	0.288
energy	594.936	0.9	0.548	0.931	0.126
drive	594.858	0.874	0.469	0.916	0.418
tooth	594.669	0.843	0.56	0.673	0.374
dress	594.342	0.884	0.668	0.925	0.148
white	591.258	0.848	0.508	0.937	0.254
fish	590.945	0.877	0.696	0.779	0.31
buy	590.893	0.92	0.536	0.593	0.339
idea	589.91	0.885	0.442	0.802	0.084
lunch	589.848	0.844	0.54	0.707	0.264
signal	588.948	0.851	0.436	0.478	0.082
shampoo	588.005	0.646	0.498	0.568	0.275
happy	586.941	0.92	0.44	0.588	0.429
stream	586.838	0.769	0.482	0.899	0.306
pay	586.778	0.944	0.544	0.609	0.183
onion	586.162	0.79	0.849	0.945	0.129
conclusion	585.476	0.803	0.419	0.813	0.03
pick	584.73	0.758	0.332	0.641	0.094
listen	583.6	0.87	0.374	0.831	0.392
love	582.612	0.929	0.458	0.514	0.163
black	582.09	0.86	0.61	0.731	0.223
bath	581.175	0.893	0.733	0.843	0.237
bowl	580.471	0.891	0.672	0.791	0.202
map	580.203	0.878	0.499	0.978	0.136
computer	579.268	0.936	0.586	0.682	0.328
public	578.421	0.919	0.407	0.95	0.167
music	576.542	0.912	0.589	0.904	0.214
couch	575.305	0.829	0.503	0.712	0.244
fact	574.174	0.808	0.369	0.608	0.138
fast	573.158	0.908	0.523	0.524	0.407
right	572.673	0.905	0.4	0.769	0.089
express	570.646	0.85	0.382	0.611	0.032
balloon	568.568	0.923	0.586	0.987	0.174
want	566.735	0.918	0.579	0.466	0.027
news	566.312	0.935	0.391	0.627	0.202
cheap	565.835	0.873	0.469	0.636	0.148

accept	1	0.33	0.85	14	0.53	0.09	0.28
stripe	1	0.61	0.47	3139	0.76	0.02	0.17
drawer	1	0.6	0.75	1215	0.9	0.12	0.17
bark	1	0.73	0.79	428	1	0.09	0.38
content	1	0.45	0.85	148	0.84	0.11	0
disease	1	0.65	0.93	361	0.86	0.03	0.17
boundary	1	0.45	0.73	210	0.81	0.04	0.3
cancer	1	0.69	0.93	373	0.76	0.03	0.21
storage	1	0.54	0.81	542	0.65	0.08	0.04
grill	1	0.84	0.84	1221	0.96	0.11	0.22
original	1	0.44	0.82	88	0.91	0.11	0.19
word	1	0.58	0.87	75	0.98	0.06	0.27
cover	1	0.37	0.82	36	0.74	0.11	0.23
status	1	0.39	0.81	28	0.92	0.12	0.17
tongue	1	0.47	0.78	55	0.74	0.11	0.41
transplant	1	0.56	0.68	365	0.47	0.02	0.12
heart	1	0.45	0.83	66	0.9	0.13	0.32
jacket	1	0.69	0.86	442	0.72	0.07	0.29
quality	1	0.45	0.92	67	0.85	0	0.14
conclusion	1	0.42	0.8	73	0.81	0.12	0.09
step	1	0.39	0.87	23	0.59	0.04	0.22
climate	1	0.52	0.92	181	0.78	0.05	0.37
development	1	0.53	0.92	227	0.63	0.06	0.21
specific	1	0.48	0.87	155	0.82	0.1	0.21
whisper	1	0.6	0.71	1730	0.62	0.09	0.26
shopping	1	0.51	0.83	38	0.9	0.06	0.38
nursery	1	0.55	0.85	153	0.56	0.05	0.31
celebration	1	0.69	0.89	202	0.79	0.1	0.27
pattern	1	0.5	0.78	302	0.99	0.15	0.07
audience	1	0.48	0.85	93	0.73	0.13	0.09
population	1	0.5	0.92	194	0.63	0.04	0.2
infinity	1	0.51	0.53	2145	0.6	0.08	0.06
delicate	1	0.45	0.75	445	0.62	0.09	0.25
idiot	1	0.59	0.8	349	0.43	0.06	0.2
lord	1	0.56	0.8	1685	0.73	0.1	0.06
executive	1	0.56	0.86	233	0.65	0.07	0.22
theme	1	0.56	0.85	141	0.76	0.1	0.18
bend	1	0.64	0.89	1087	0.83	0.04	0.29
squeeze	1	0.73	0.72	1536	0.81	0.11	0.16
pity	1	0.44	0.78	66	0.57	0.08	0.26
count	1	0.41	0.84	27	0.92	0.12	0.07
store	1	0.55	0.83	65	0.93	0.05	0.5
touch	1	0.39	0.89	17	0.93	0.05	0.29
pound	1	0.57	0.79	216	0.72	0.13	0.24
cabbage	1	0.84	0.67	3264	0.9	0.11	0.36
lose	1	0.42	0.89	26	0.63	0.05	0.21
mixture	1	0.63	0.77	314	0.99	0.07	0.1
antique	1	0.72	0.82	1120	0.7	0.07	0.18
fraud	1	0.59	0.64	533	0.84	0.12	0.08
similar	1	0.4	0.89	20	0.91	0.04	0.21
war	1	0.55	0.89	308	0.56	0.09	0.38
joke	1	0.52	0.84	222	0.74	0.1	0.49
noisy	1	0.6	0.88	454	0.42	0.07	0.23
deceive	1	0.56	0.75	638	0.4	0.06	0.1
list	1	0.36	0.82	35	0.75	0.11	0.14
destruction	1	0.66	0.88	790	0.97	0.09	0.4
amusing	1	0.54	0.81	426	0.43	0.05	0.33
train	1	0.4	0.86	25	0.66	0.08	0.36
total	1	0.48	0.87	97	0.6	0.05	0.13
coin	1	0.5	0.73	311	0.44	0.08	0.36
response	1	0.46	0.94	21	0.93	0.01	0.09
insight	1	0.46	0.73	151	0.71	0.06	0.09
text	1	0.49	0.87	94	0.69	0.06	0.09
secret	1	0.48	0.85	22	0.81	0.05	0.06
treasure	1	0.5	0.82	529	0.52	0.06	0.23
death	1	0.54	0.83	576	0.69	0.11	0.36
guide	1	0.38	0.86	19	0.7	0.05	0.06
sugar	1	0.71	0.85	1079	0.82	0.09	0.21
pencil	1	0.6	0.87	529	0.5	0.06	0.16

boat	565.338	0.927	0.601	0.773	0.187
join	564.655	0.892	0.467	0.533	0.157
protein	564.475	0.674	0.588	0.585	0.291
kid	562.831	0.939	0.537	0.936	0.087
agree	561.185	0.89	0.37	0.741	0.27
yellow	554.798	0.835	0.656	0.996	0.29
baby	553.38	0.905	0.556	0.897	0.211
court	551.058	0.837	0.478	0.729	0.453
final	540.662	0.872	0.457	0.858	0.091

current	1	0.4	0.87	20	0.49	0.07	0.06
break	1	0.46	0.88	13	0.44	0.04	0.18
join	1	0.47	0.89	37	0.53	0.08	0.32
kidney	1	0.61	0.68	584	0.67	0.03	0.21
alternative	1	0.47	0.9	75	0.71	0.03	0.17
nation	1	0.54	0.87	241	0.95	0.12	0.1
cheese	1	0.84	0.86	1702	0.94	0.1	0.37
playground	1	0.56	0.84	354	0.52	0.09	0.23
nose	1	0.7	0.81	847	0.93	0.1	0.33

Appendix L: L2 and L1 (ELP) Model Comparisons Statistics

L2 Independent Model Comparisons

	Model Description			***Test Against Prior Model***	
Mo-Del	***Fixed Effects***	***Random Effects***	***AIC***	***Statistic***	***p***
1	None	language	1219.7		
2	Model 1 + L2 Word Naming RT	language	1146.7	$X^2(1) = 74.975$	<.005
3	Model 2 + L2 Word Naming Accuracy	language	1147.9	$X^2(1) = 0.373$	0.37

L2 Integrated Model Comparisons

		Model Description		***Test Against Prior Model***	
Mo-del	***Fixed Effects***	***Random Effects***	***AIC***	***Statistic***	***p***
1	None	Language	1556.2		
2	Model 1 - L2 Word Naming RT	Language	1547.0	$X^2(1) = 11.138$	<.005
3	Model 2 - L2 Word Naming Accuracy	Language	1538.5	$X^2(1) = 10.467$	<.005

L1 Independent Model Comparisons

	Model Description			***Test Against Prior Model***	
Mo-del	***Fixed Effects***	***Random Effects***	***AIC***	***Statistic***	***p***
1	None	language	1219.7		
2	Model 2 + ELP Word Naming RT	language	1162.1	$X^2(1) = 59.638$	<.005
3	Model 3 + ELP Word Naming SD	language	1164.0	$X^2(1) = 0.0346$	0.85
4	Model 3 + ELP Word Naming Accuracy	language	1184.9	$X^2(1) = 0.0164$	0.9

L1 Integrated Model Comparisons

	Model description			***Test against prior model***	
Model	***Fixed Effects***	***Random Effects***	***AIC***	***Statistic***	***p***
1	None	language	1556.2		
2	Model 2 + ELP Word Naming RT	language	1537.0	$X^2(1) = 21.119$	<.005
3	Model 3 + ELP Word Naming Accuracy	language	1530.9	X2(1) = 8.1724	<.005

Combined Independent Model Comparisons

	Model description			***Test against prior model***	
Model	***Fixed Effects***	***Random Effects***	***AIC***	***Statistic***	***p***
1	None	language	1219.7		
2	Model 1 + L2 Word Naming RT	language	1146.7	$X^2(1) = 59.638$	<.005
3	Model 2 + ELP Word Naming RT	language	1140.0	$X^2(1) = 8.7091$	<.005
4	Model 3 + L2 Word Naming Accuracy	language	1141.5	$X^2(1) = 0.5527$	0.45
5	Model 3 + ELP Word Naming SD	language	1141.9	$X^2(1) = 0.1233$	0.73
6	Model 3 + ELP Word Naming Accuracy	language	1142.0	$X^2(1) = 0.0352$	0.85

Combined Integrated Model Comparisons

	Model Description			***Test Against Prior Model***	
Model	***Fixed Effects***	***Random Effects***	***AIC***	***Statistic***	***p***
1	None	language	1556.2		
2	Model 1 + ELP Word Naming RT	language	1537.0	$X^2(1) = 21.119$	<.005
3	Model 2 + L2 Word Naming Accuracy	language	1530.7	$X^2(1) = 8.3754$	<.005
4	Model 3 + L2 Word Naming RT Mean Score	language	1532.7	$X^2(1) = 0.0069$	0.93
5	Model 3 + ELP Accuracy -	language	1551.5	$X^2(1) = 6.207$	0.01

Appendix M: Individual Output with Word Recognition Indices – Independent Task

Independent essay – Score 5

I think it is more important to choose study subjects that you are interested in, rather than to choose subjects that prepare you for a certain career. Choosing subjects is an important decision that people make because

it affects their future. You do not want to choose a subject and then later down the line realize that this is not what you really want to do. In that case, you might find yourself confused in a midlife crisis.

It is important to choose to do what you are really interested in. Firstly, if you are doing something that interests you, then you will enjoy doing it. It is only when you enjoy doing something that you can fully use your potential and do your best in it. It will provide you with comfort and satisfaction in life. For example if you like your job, then you can really excell in it and make a good career.

Secondly if you are interested in something, you are more likely to want to stay in that field for the most part of your life or career. If you choose subjects that you are not really interested then at one point you will start to get distracted. You will start finding the work that you do tedious and you may not enjoy it. That will hamper your success and your progress. But most of all you may not be happy with what you are doing. That takes away your ability to fully utilise your potential.

To illustrate how important it is to choose subjects of interest many examples can be used. Say that Joe is a student in high school. He has always thought of being an engineer when he grows up. He takes all the science and mathematics subjects and goes to an engineering school where he studies mechanical engineer. After college he inturns at an engineering firm and later goes on to do masters in fluid mechanics. But what Joe didn't think about is that his real interest lies in economics. It fascinates him to think about how people interact with the economy and how it all works. Slowly he starts getting distracted from his work at the engineering firm. He finds the long calculations tedious and boring. Working with the huge complicated systems give him a headache and he realizes that this is not what he really wanted to be. What does Joe do now? He has made a lot of progress in the field and if he wishes to switch now, it will mean that he has to start over again.

Therefore, I believe that choosing subjects of interest is most important because in order to succeed and lead a satisfactory like you must pursue what really interests you. However you must really think about what interests you the most. It helps sometimes to have a certain goal or vision in mind that you work up to. Otherwise you mind find yourself getting distracted and deviating from the good way of life.

Independent essay – Score 1.5

in my opinion .. well, im not fron this country but i think that the world is now thinking more of other people i think that the peole now are looking fowere for other people because we got a new generation that is coming up to the hill that is not good full of bad things one of does thing is drugs and alcohol ,litter kids now drink and smoke litter kids now kill and even care this new generation is going to take care of this world in the badess way they can , for us the good people is very bad because we dont whant are kids to be like that and if we dont take care of other people the world is going to keep going the same track until it reck . when that happend we wont be able to do nothis just to see what more is going to happend in this world .

if we dont take care of this country as quikli as we can we are going to lose control of it and the bad people is going to take ccare of it in any secand , but we cant let then do that when we still here washing then grow as they want . we can stop this this is not someting impossible if we whant we can do it we stell have the power the only thing that we need to do is handle of it and stay with the power to have a better world and to a better life .

the way to do this is taking care of the people that live in the street or the people that realy needs help with there problems becuse that is the vasic thind the problems and if we can solve this this well be a better world

Scores	***L2 RT***	***ELP RT***	***Overlapping ELP SD***	***Overlapping ELP Accuracy***
5	647.278	612.728	128.328	0.989
1.5	616.558	596.805	116.768	0.997

High-scored essay

Words	*Token count*	*L2 RT*	*Words*	*Token count*	*ELP RT*	*Words (types)*	*Token count*	*Overlapping ELP SD*	*Words (types)*	*Token count*	*Overlapping ELP Accuracy*
pursue	1	842	deviating	1	813	make	2	293	ability	1	1
engineering	3	838	interests	3	720	point	1	266	again	1	1
mechanical	1	833	hamper	1	713	use	1	247	always	1	1
therefore	1	832	success	1	713	pursue	1	245	away	1	1
does	1	772	pursue	1	711	again	1	236	be	3	1
thought	1	768	economics	1	707	confused	1	227	being	1	1
crisis	1	767	tedious	2	705	firm	2	224	believe	1	1
interested	4	767	mechanical	1	697	lead	1	213	boring	1	1
realizes	1	767	economy	1	694	future	1	212	can	3	1
certain	2	765	student	1	694	interest	3	211	case	1	1
complicated	1	765	succeed	1	693	good	2	205	certain	2	1
their	1	765	certain	2	692	success	1	202	choosing	2	1
college	1	760	engineering	3	691	engineering	3	199	comfort	1	1
potential	2	759	mechanics	1	690	does	1	197	complicated	1	1
examples	1	757	future	1	689	enjoy	3	196	confused	1	1
prepare	1	752	school	2	678	important	5	195	crisis	1	1
fluid	1	746	calculations	1	676	mind	2	187	decision	1	1
then	4	745	decision	1	675	economics	1	184	do	7	1
career	3	735	important	5	674	succeed	1	181	doing	4	1
doing	4	732	complicated	1	670	realize	1	177	engineer	2	1
wishes	1	725	interested	4	669	certain	2	175	enjoy	3	1
engineer	2	724	confused	1	667	mechanical	1	175	example	1	1
headache	1	724	crisis	1	666	goes	2	172	examples	1	1
likely	1	719	choose	6	665	working	1	172	field	2	1
economics	1	709	make	2	665	college	1	169	find	2	1
realize	1	708	choosing	2	663	engineer	2	167	finds	1	1
systems	1	708	secondly	1	662	be	3	165	firm	2	1
boring	1	695	study	1	662	economy	1	164	fluid	1	1
goes	2	693	distracted	3	661	doing	4	162	future	1	1
huge	1	693	satisfaction	1	661	provide	1	159	give	1	1
progress	2	692	their	1	660	goal	1	155	goal	1	1
choosing	2	690	satisfactory	1	659	fluid	1	153	goes	2	1
economy	1	690	firm	2	657	only	1	152	good	2	1
confused	1	688	systems	1	657	choose	6	151	happy	1	1
affects	1	687	starts	1	654	therefore	1	151	has	3	1
field	2	681	college	1	653	have	1	150	have	1	1
might	1	681	really	7	651	realizes	1	150	headache	1	1
where	1	679	examples	1	650	progress	2	147	helps	1	1
people	2	678	fascinates	1	649	comfort	1	147	high	1	1
wanted	1	675	realize	1	649	way	1	147	job	1	1
later	2	674	stay	1	649	school	2	145	later	2	1
finds	1	673	studies	1	646	life	3	141	lies	1	1
succeed	1	669	fully	2	645	later	2	141	life	3	1
subjects	7	666	engineer	2	644	crisis	1	137	likely	1	1
vision	1	665	illustrate	1	642	works	1	136	line	1	1
being	1	664	interact	1	642	subject	1	136	long	1	1
be	3	663	subject	1	642	prepare	1	134	made	1	1
interest	3	662	however	1	641	interested	4	132	make	2	1
masters	1	662	slowly	1	641	complicated	1	131	masters	1	1
comfort	1	659	being	1	640	line	1	131	may	2	1
provide	1	653	point	1	640	sometimes	1	130	mean	1	1
sometimes	1	653	fluid	1	639	stay	1	130	mechanical	1	1
mean	1	650	give	1	639	decision	1	128	must	2	1
has	3	649	progress	2	636	systems	1	128	not	6	1
rather	1	649	comfort	1	635	potential	2	126	now	2	1
lies	1	648	therefore	1	635	case	1	126	one	1	1
believe	1	647	case	1	634	takes	2	123	only	1	1
ability	1	645	good	2	634	get	1	123	order	1	1
made	1	645	provide	1	634	believe	1	123	part	1	1
works	1	645	realizes	1	632	example	1	121	people	2	1

science	1	643	switch	1	632	wanted	1	120	point	1	1
real	1	642	goal	1	631	wishes	1	120	potential	2	1
must	2	637	enjoy	3	630	career	3	119	prepare	1	1
student	1	637	interest	3	630	student	1	119	provide	1	1
firm	2	636	be	3	627	now	2	119	pursue	1	1
important	5	636	can	3	627	give	1	119	rather	1	1
think	4	636	say	1	627	where	1	119	real	1	1
always	1	634	mathematics	1	625	examples	1	118	realize	1	1
working	1	634	does	1	624	people	2	117	realizes	1	1
want	3	628	start	3	624	helps	1	116	say	1	1
subject	1	627	finds	1	623	huge	1	115	school	2	1
part	1	622	career	3	621	slowly	1	115	science	1	1
enjoy	3	619	getting	2	621	high	1	114	slowly	1	1
give	1	619	subjects	7	619	then	4	114	sometimes	1	1
work	3	619	use	1	618	affects	1	112	start	3	1
have	1	616	headache	1	617	being	1	111	stay	1	1
decision	1	611	potential	2	617	boring	1	107	student	1	1
takes	2	611	mind	2	616	how	3	107	study	1	1
order	1	609	line	1	615	choosing	2	105	subject	1	1
away	1	608	example	1	614	always	1	103	succeed	1	1
find	2	608	prepare	1	612	real	1	102	success	1	1
how	3	608	rather	1	611	has	3	102	switch	1	1
example	1	605	doing	4	610	vision	1	101	systems	1	1
not	6	605	science	1	609	one	1	100	takes	2	1
choose	6	604	huge	1	608	masters	1	99	their	1	1
job	1	604	may	2	606	mean	1	99	therefore	1	1
happy	1	603	again	1	606	finds	1	98	think	4	1
future	1	602	field	2	606	lies	1	98	use	1	1
make	2	601	affects	1	605	happy	1	97	used	1	1
way	1	600	grows	1	605	can	3	97	vision	1	1
helps	1	596	lead	1	605	their	1	97	want	3	1
may	2	594	sometime	1	605	long	1	96	wanted	1	1
do	7	592	then	4	602	might	1	96	way	1	1
good	2	592	takes	2	601	want	3	94	where	1	1
mind	2	591	get	1	600	rather	1	93	will	6	1
switch	1	589	thought	1	599	made	1	93	wishes	1	1
one	1	585	goes	2	598	away	1	92	work	3	1
now	2	581	boring	1	596	start	3	92	working	1	1
lead	1	580	vision	1	594	may	2	88	works	1	1
goal	1	579	believe	1	592	study	1	84	might	1	0.966
slowly	1	579	have	1	591	will	6	84	career	3	0.966
case	1	578	later	2	591	likely	1	83	college	1	0.964
start	3	578	works	1	591	job	1	83	does	1	0.964
only	1	576	are	7	590	think	4	83	economics	1	0.964
used	1	574	has	3	587	science	1	82	how	3	0.964
success	1	573	where	1	587	ability	1	82	lot	1	0.964
get	1	561	real	1	586	must	2	82	affects	1	0.963
line	1	560	working	1	586	say	1	80	economy	1	0.963
life	3	558	ability	1	583	part	1	79	lead	1	0.963
will	6	557	do	7	583	work	3	78	subjects	7	0.963
lot	1	556	high	1	583	subjects	7	75	then	4	0.963
high	1	555	finding	1	582	field	2	75	engineering	3	0.962
stay	1	554	is	10	582	order	1	75	important	5	0.962
again	1	551	find	2	580	not	6	75	mind	2	0.962
say	1	547	want	3	580	find	2	74	progress	2	0.962
use	1	546	way	1	580	switch	1	74	thought	1	0.962
can	3	541	wanted	1	579	do	7	72	interested	4	0.96
study	1	541	now	2	578	thought	1	71	interest	3	0.957
school	2	540	made	1	577	headache	1	68	huge	1	0.929
point	1	539	people	2	574	used	1	59	choose	6	0.926
long	1	529	joe	3	573	lot	1	58	get	1	0.923
			only	1	573						
			used	1	573						
			one	1	571						
			think	4	571						
			helps	1	570						
			didn't	1	569						
			lies	1	569						
			masters	1	569						

best	1	568
your	8	568
long	1	567
wishes	1	566
likely	1	565
not	6	565
must	2	561
job	1	558
will	6	558
happy	1	556
might	1	555
part	1	555
life	3	553
order	1	549
how	3	547
away	1	545
mean	1	545

Low-scored essay

Words	*Token count*	*L2 RT*	*Words*	*Token count*	*ELP RT*	*Words (types)*	*Token count*	*Overlapping ELP SD*	*Words (types)*	*Token count*	*Overlapping ELP Accuracy*
does	1	772	generation	2	699	things	1	272	able	1	1
then	2	745	control	1	687	new	2	240	alcohol	1	1
thing	2	724	still	1	684	even	1	222	bad	3	1
needs	1	720	needs	1	672	good	2	205	be	3	1
things	1	714	things	1	670	able	1	198	can	5	1
taking	1	713	street	1	655	does	1	197	care	5	1
drugs	1	688	stay	1	649	help	1	182	control	1	1
problems	2	686	see	1	644	needs	1	179	country	2	1
impossible	1	681	stop	1	642	full	1	174	do	5	1
looking	1	681	same	1	641	be	3	165	drink	1	1
people	7	678	full	1	638	grow	1	164	drugs	1	1
generation	2	664	taking	1	635	generation	2	159	even	1	1
be	3	663	good	2	634	just	1	159	full	1	1
opinion	1	658	track	1	634	only	1	152	generation	2	1
here	1	653	new	2	632	taking	1	152	going	6	1
thinking	1	651	country	2	630	have	2	150	good	2	1
litter	2	650	coming	1	629	country	2	149	grow	1	1
even	1	636	able	1	628	way	2	147	handle	1	1
live	1	636	be	3	627	life	1	141	have	2	1
think	2	636	can	5	627	track	1	130	help	1	1
want	1	628	thinking	1	626	stay	1	130	here	1	1
my	1	625	even	1	625	control	1	124	hill	1	1
lose	1	623	does	1	624	opinion	1	123	impossible	1	1
world	6	622	solve	1	624	impossible	1	123	keep	1	1
well	2	621	impossible	1	620	very	1	120	kill	1	1
going	6	619	wont	1	619	lose	1	120	let	1	1
very	1	618	opinion	1	618	bad	3	119	life	1	1
country	2	616	kids	3	615	now	4	119	litter	2	1
have	2	616	thing	2	615	stop	1	119	looking	1	1
alcohol	1	608	problems	2	613	people	7	117	my	1	1
track	1	608	smoke	1	612	litter	2	117	need	1	1
solve	1	607	very	1	609	hill	1	117	needs	1	1
still	1	607	lose	1	607	then	2	114	new	2	1
not	3	605	then	2	602	going	6	114	not	3	1
help	1	604	washing	1	602	alcohol	1	113	now	4	1
street	1	602	going	6	601	thinking	1	111	one	1	1
care	5	601	handle	1	601	problems	2	111	only	1	1
power	2	600	got	1	595	solve	1	107	opinion	1	1
way	2	600	have	2	591	street	1	107	people	7	1
hill	1	599	are	3	590	care	5	106	power	2	1
kill	1	595	alcohol	1	586	same	1	105	problems	2	1
do	5	592	keep	1	586	looking	1	103	same	1	1

good	2	592	litter	2	586	take	4	100	see	1	1
see	1	589	take	4	585	one	1	100	smoke	1	1
just	1	587	do	5	583	handle	1	98	solve	1	1
full	1	585	is	13	582	can	5	97	stay	1	1
one	1	585	want	1	580	want	1	94	still	1	1
able	1	584	way	2	580	my	1	94	stop	1	1
grow	1	584	bad	3	579	thing	2	93	street	1	1
need	1	583	my	1	578	drugs	1	91	take	4	1
new	2	582	now	4	578	smoke	1	91	thing	2	1
same	1	582	care	5	576	still	1	90	things	1	1
now	4	581	help	1	576	well	2	87	think	2	1
take	4	579	grow	1	575	drink	1	85	track	1	1
handle	1	578	here	1	575	see	1	84	very	1	1
only	1	576	just	1	574	keep	1	83	want	1	1
let	1	571	people	7	574	think	2	83	way	2	1
drink	1	570	only	1	573	power	2	80	well	2	1
control	1	560	better	3	572	here	1	79	world	6	1
life	1	558	one	1	571	live	1	78	taking	1	0.966
bad	3	554	think	2	571	not	3	75	thinking	1	0.966
stay	1	554	looking	1	567	let	1	73	does	1	0.964
smoke	1	553	not	3	565	do	5	72	just	1	0.964
can	5	541	power	2	565	need	1	71	live	1	0.963
stop	1	517	need	1	563	kill	1	71	then	2	0.963
keep	1	501	drink	1	560	world	6	61	lose	1	0.88
			world	6	559						
			kill	1	557						
			well	2	555						
			live	1	554						
			life	1	553						
			drugs	1	552						
			hill	1	551						
			let	1	530						

Appendix N: Individual Output with Word Recognition Indices – Integrated Task

Integrated essay – Score = 4.5

The lecturer rebutts some of the points made in the reading passage by challenging their assertions. Firstly, the lecture states that wild fish are already endengered and thus the risk of spreading disease and infection are minimal. Whilst the reading passage highlights the huge risk of disease and infection, the lecturer states the positive by stating that fish farming gives wild fish an opportunity to rebound and accumulate in numbers. Thus the lecturer emphasizes the role fish farming plays in combating endengerment.

Secondly, the lecturer also downplays the health risk humans face when consuming chemicaly treated farm fish. The lecturer compares poultry and livestock that have undergone growth inducing chemicals and asserts that no known harm has come from consuming them. He further points out that fish feed with growth-inducing chemicals have a better nutrional value than wild fish. This challenges the reading passage assertion that people can be exposed to harmful or unatural long term effects from consuming farm raised fish.

Last but not least the lecturer also claims that the species used to feed the farm raised fish are not edible by humans, and thus the premise stated in the reading passage that protein is being reduced from the sea is false. This notion also rejects the premise of long term wastfulness.

Integrated essay – Score = 1

Over fourty years fish farming has grown near the shoreline.

Fact of the fish farming is, that it became an increasingley common method of the production from fish, the fact of the fish farming is that today almost one third of the fish consumed are grown on these farms.

The fish farming is a huge busniss, but it brings a lot of different problems with it.

So there are reasons against the fish farming.

It is for sure that cause of the fish farming the healthy in wildness living fish can get very ill, reasons for the ilness are being in small areas like the enclosures in farmings.

But for the ilness farmers can do something about it they can use medecines or to help their own fishes. Human also can get very sick from eating the fish because the farmers want to make a lot of money and they want to do quick so they over feed the fishes and save money on the food they feed to the fishes these are reasons why human could get sick.

Scores	L2 RT	L2 Accuracy	ELP RT	ELP Accuracy	Overlapping ELP SD	Overlapping ELP Accuracy
4.5	654.521	0.946	631.684	0.989	137.409	0.9909
1	624.472	0.986	612.318	0.987	119.950	0.9899

High-scored essay

Word (types)	Token count	L2 accuracy	Word (types)	Token count	ELP accuracy	Word (types)	Token count	L2 RT	Word (types)	Token count	Overlapping ELP SD	Word (types)	Token count	Overlapping ELP Accuracy
premise	2	0.556	accumulate	1	1	premise	2	860	humans	2	328.2	long	2	1
species	1	0.76	already	1	1	spreading	1	789	sea	1	281.8	can	1	1
wild	3	0.806	also	3	1	accumulate	1	773	opportunity	1	274.5	sea	1	1
lecturer	6	0.81	asserts	1	1	stated	1	771	feed	2	241.5	out	1	1
treated	1	0.815	be	1	1	their	1	765	premise	2	212.4	no	1	1
stated	1	0.821	being	1	1	claims	1	743	accumulate	1	200.7	face	1	1
have	2	0.852	better	1	1	treated	1	742	passage	4	197.5	also	3	1
protein	1	0.862	can	1	1	effects	1	721	last	1	175.6	used	1	1
value	1	0.889	challenges	1	1	humans	2	719	states	2	172.2	harm	1	1
accumulate	1	0.917	challenging	1	1	passage	4	717	lecturer	6	170.5	fish	9	1
being	1	0.923	chemicals	2	1	species	1	716	be	1	164.6	risk	3	1
raised	2	0.929	claims	1	1	highlights	1	712	plays	1	160.2	last	1	1
claims	1	0.931	compares	1	1	lecturer	6	707	disease	2	158.2	not	2	1
known	1	0.933	consuming	3	1	disease	2	700	lecture	1	151.7	plays	1	1
notion	1	0.935	disease	2	1	least	1	696	have	2	150.1	have	2	1
harm	1	0.962	edible	1	1	states	2	695	known	1	150.0	farm	3	1
spreading	1	0.962	effects	1	1	notion	1	693	effects	1	147.1	opportunity	1	1
can	1	0.963	face	1	1	huge	1	693	farm	3	145.4	wild	3	1
effects	1	0.963	farm	3	1	positive	1	691	risk	3	142.8	made	1	1
passage	4	0.963	farming	2	1	points	2	686	protein	1	140.5	has	1	1
disease	2	0.964	fish	9	1	protein	1	684	species	1	139.9	be	1	1
highlights	1	0.964	further	1	1	lecture	1	683	positive	1	137.6	growth	1	1
least	1	0.964	gives	1	1	people	1	678	notion	1	130.1	being	1	1
has	1	0.966	growth	1	1	known	1	674	reading	4	127.5	numbers	1	1
huge	1	0.966	harm	1	1	raised	2	674	wild	3	127.2	known	1	1
numbers	1	0.966	harmful	1	1	numbers	1	673	no	1	126.0	raised	2	1
points	2	0.966	has	1	1	being	1	664	treated	1	125.4	people	1	1
also	3	1	have	2	1	growth	1	664	points	2	119.0	protein	1	1
be	1	1	highlights	1	1	be	1	663	people	1	117.1	points	2	1
come	1	1	humans	2	1	value	1	653	out	1	115.4	positive	1	1
face	1	1	inducing	1	1	has	1	649	huge	1	115.3	notion	1	1
farm	3	1	infection	2	1	made	1	645	health	1	112.8	states	2	1
feed	2	1	is	2	1	wild	3	645	being	1	111.4	least	1	1
fish	9	1	known	1	1	feed	2	637	stated	1	110.9	disease	2	1
growth	1	1	least	1	1	opportunity	1	634	highlights	1	104.5	lecturer	6	1
health	1	1	lecturer	6	1	farm	3	628	has	1	102.3	highlights	1	1
humans	2	1	livestock	1	1	health	1	623	harm	1	100.8	species	1	1
last	1	1	long	2	1	come	1	622	can	1	97.1	passage	4	1
lecture	1	1	made	1	1	reading	4	621	spreading	1	96.9	humans	2	1
long	2	1	minimal	1	1	have	2	616	their	1	96.7	effects	1	1
made	1	1	no	1	1	plays	1	610	long	2	96.0	treated	1	1

no	1	1	not	2	1	not	2	605	value	1	95.9	claims	1	1
not	2	1	notion	1	1	last	1	600	fish	9	93.8	their	1	1
opportunity	1	1	numbers	1	1	risk	3	600	made	1	92.7	stated	1	1
out	1	1	opportunity	1	1	fish	9	594	least	1	91.0	accumulate	1	1
people	1	1	out	1	1	harm	1	576	come	1	89.8	spreading	1	1
plays	1	1	passage	4	1	used	1	574	claims	1	88.5	lecture	1	0.964
positive	1	1	people	1	1	also	3	573	raised	2	88.0	come	1	0.963
reading	4	1	plays	1	1	face	1	569	growth	1	86.1	feed	2	0.963
risk	3	1	points	2	1	no	1	560	also	3	85.8	reading	4	0.962
sea	1	1	positive	1	1	out	1	555	face	1	82.8	huge	1	0.929
states	2	1	protein	1	1	sea	1	544	numbers	1	79.9	health	1	0.926
their	1	1	raised	2	1	can	1	541	not	2	74.6	value	1	0.926
used	1	1	reduced	1	1	long	2	529	used	1	59.2	premise	2	0.846
			risk	3	1									
			role	1	1									
			sea	1	1									
			species	1	1									
			spreading	1	1									
			stated	1	1									
			states	2	1									
			stating	1	1									
			term	2	1									
			their	1	1									
			thus	3	1									
			treated	1	1									
			undergone	1	1									
			used	1	1									
			wild	3	1									
			combating	1	0.96									
			come	1	0.96									
			emphasizes	1	0.96									
			exposed	1	0.96									
			feed	2	0.96									
			lecture	1	0.96									
			poultry	1	0.96									
			reading	4	0.96									
			rebound	1	0.96									
			rejects	1	0.96									
			FALSE	1	0.96									
			are	3	0.93									
			health	1	0.93									
			huge	1	0.93									
			value	1	0.93									
			assertions	1	0.89									
			premise	2	0.85									
			assertion	1	0.84									
			whilst	1	0.78									

Low-scored essay

Word (types)	*Token count*	*L2 accuracy*	*Word (types)*	*Token count*	*L2 RT*	*Word (types)*	*Token count*	*ELP accuracy*	*Word (types)*	*Token count*	*Overlapping ELP SD*	*Word (types)*	*Token count*	*Overlapping ELP Accuracy*
grown	2	0.893	farmers	2	812	almost	1	1	make	1	292.9	can	4	1
being	1	0.923	became	1	768	also	1	1	almost	1	290.1	small	1	1
ill	1	0.958	their	1	765	areas	1	1	use	1	247.2	use	1	1
reasons	3	0.960	method	1	735	became	1	1	feed	2	241.5	common	1	1
sick	2	0.960	third	1	729	being	1	1	living	1	240.5	also	1	1
production	1	0.962	areas	1	711	brings	1	1	farmers	2	193.3	one	1	1
small	1	0.962	huge	1	693	can	4	1	help	1	182.5	do	2	1
can	4	0.963	grown	2	693	cause	1	1	small	1	173.4	fish	1	1
healthy	1	0.963	problems	1	686	common	1	1	healthy	1	169.0	make	1	1
became	1	0.964	different	1	677	different	1	1	areas	1	163.5	help	1	1
money	2	0.964	years	1	671	do	2	1	production	1	150.3	why	1	1
own	1	0.964	eating	1	666	eating	1	1	quick	1	138.6	almost	1	1

has	1	0.966	reasons	3	665	enclosures	1	1	common	1	137.2	food	1	1
huge	1	0.966	being	1	664	fact	2	1	became	1	127.8	quick	1	1
method	1	0.967	healthy	1	655	farmers	2	1	money	2	126.2	very	2	1
almost	1	1	ill	1	649	farming	6	1	get	3	122.8	cause	1	1
also	1	1	has	1	649	farms	1	1	different	1	122.1	fact	2	1
areas	1	1	money	2	644	fish	1	1	very	2	120.3	want	2	1
cause	1	1	production	1	641	food	1	1	eating	1	117.5	living	1	1
common	1	1	feed	2	637	has	1	1	fact	2	117.3	human	2	1
different	1	1	own	1	635	healthy	1	1	grown	2	115.8	own	1	1
do	2	1	human	2	633	help	1	1	sick	2	115.5	production	1	1
eating	1	1	living	1	629	human	2	1	huge	1	115.3	money	2	1
fact	2	1	want	2	628	ill	1	1	years	1	111.5	has	1	1
farmers	2	1	fact	2	625	is	4	1	being	1	111.4	ill	1	1
feed	2	1	cause	1	618	living	1	1	problems	1	110.6	healthy	1	1
fish	1	1	very	2	618	make	1	1	has	1	102.3	being	1	1
food	1	1	quick	1	614	method	1	1	third	1	101.7	reasons	3	1
get	3	1	food	1	610	money	2	1	one	1	100.0	eating	1	1
help	1	1	today	1	607	one	1	1	can	4	97.1	years	1	1
human	2	1	almost	1	606	problems	1	1	their	1	96.7	different	1	1
living	1	1	why	1	605	production	1	1	today	1	94.3	problems	1	1
lot	2	1	help	1	604	quick	1	1	want	2	94.0	areas	1	1
make	1	1	make	1	601	reasons	3	1	fish	1	93.8	third	1	1
one	1	1	fish	1	594	shoreline	1	1	method	1	91.9	method	1	1
problems	1	1	do	2	592	small	1	1	human	2	90.2	their	1	1
quick	1	1	one	1	585	their	1	1	own	1	86.9	became	1	1
sure	1	1	also	1	573	these	2	1	also	1	85.8	farmers	2	1
their	1	1	common	1	568	third	1	1	food	1	85.6	sure	1	0.966
third	1	1	get	3	561	use	1	1	reasons	3	83.6	sick	2	0.964
today	1	1	lot	2	556	very	2	1	sure	1	82.9	lot	2	0.964
use	1	1	sure	1	554	want	2	1	do	2	72.4	today	1	0.963
very	2	1	use	1	546	why	1	1	cause	1	72.2	feed	2	0.963
want	2	1	small	1	542	years	1	1	ill	1	62.2	huge	1	0.929
why	1	1	can	4	541	sure	1	0.97	lot	2	58.1	grown	2	0.926
years	1	1	sick	2	535	could	1	0.96	why	1	58.0	get	3	0.923
						feed	2	0.96						
						fishes	3	0.96						
						lot	2	0.96						
						sick	2	0.96						
						today	1	0.96						
						consumed	1	0.95						
						are	4	0.93						
						grown	2	0.93						
						huge	1	0.93						
						get	3	0.92						
						wildness	1	0.89						

Appendix O: Overlapping Independent and Integrated Model Statistics

Correlations between Essay Scores and the Overlapping ELP indices

Overlapping ELP Indices	***Independent***	***Integrated***
Overlapping ELP Word Naming SD mean	0.186^{***}	0.170^{***}
Overlapping ELP Word Naming Reaction Time Mean	0.132^{***}	0.160^{***}
Overlapping ELP Word Naming Accuracy Mean	0.037	0.178^{***}

*** $p < .0005$, ** $p < .005$, * $p < 0.05$, $p > .05$

Overlapping ELP Independent Model Comparisons

Model description				Test against prior model	
Model	***Fixed Effects***	***Random Effects***	***AIC***	***Statistic***	***p***
1	None	language	1219.7		
2	Model 1 + Overlapping ELP Word Naming SD mean	language	1213.2	$X^2(1) = 8.556$	<.005
3	Model 2 + Overlapping ELP Word Naming Reaction Time Mean	language	1212.8	$X^2(1) = 2.369$	0.12
4	Model 2 + Overlapping ELP Word Naming Accuracy Mean	language	1214.1	$X^2(1) = 1.092$	0.30

Overlapping ELP Independent Model with Best Fit

Random effects	***Variance***	***SD***
Language (intercept)	0.107	0.327
Residual	0.676	0.822

Fixed effects	***Estimates***	***SE***	***t-value***	***p***	***R^2***	***95% CI***	
(Intercept)	-14.107	14.550	-0.970	0.33	0.018	0.052	0.004
Overlapping ELP Word Naming SD mean	0.019	0.007	2.854	<.005	0.016	0.045	0.001
Overlapping ELP Word Naming Accuracy Mean	15.366	14.687	1.046	0.30	0.002	0.018	0.000

Overlapping ELP Integrated Model Comparisons

Model description				Test against prior model	
Model	***Fixed Effects***	***Random Effects***	***AIC***	***Statistic***	***p***
1	None	language	1556.2		
2	Model 1 + Overlapping ELP Word Naming Accuracy Mean	language	1547.5	$X^2(1) = 10.616$	<.005
3	Model 2 + Overlapping ELP Word Naming SD mean	language	1538.2	$X^2(1) = 11.367$	<.005
4	Model 3 + Overlapping ELP Word Naming Reaction Time Mean	language	1538.0	$X^2(1) = 2.211$	0.14

Overlapping ELP Integrated Model with Best Fit

Random effects	***Variance***	***SD***					
Language (intercept)	0.111	0.334					
Residual	1.360	1.166					
Fixed effects	***Estimates***	***SE***	***t-value***	***p***	***R2***	***95% CI***	
(Intercept)	-93.253	25.604	-3.642	<.005	0.045	0.09	0.018
Overlapping ELP Word Naming Accuracy Mean	93.227	25.622	3.639	<.005	0.027	0.06	0.006
Overlapping ELP Word Naming SD mean	0.0292	0.009	3.394	<.005	0.023	0.05	0.004