Corpus approaches to issues in second language acquisition: three studies

Randy Appel

A Thesis

In the Department

of

Education

Presented in Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy (Education) at

Concordia University

Montreal, Quebec, Canada

February 2017

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Randy F. Appel

Entitled: Corpus Approaches to Issues in Second Language Acquisition: Three Studies

and submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY (Education)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____	Chair
Dr. Joanna White	
_____	External Examiner
Dr. Bill Crawford	
_____	External to Program
Dr. Norman Segalowitz	
_____	Examiner
Dr. Walcir Cardoso	
_____	Examiner
Dr. Marlise Horst	
_____	Thesis Supervisors
Dr. Pavel Trofimovich & Kim McDonough	

Approved by _____

Dr. Richard Schmid, Department Chair

Feb 6th, 2017 _____

Dr. André Roy, Dean of Faculty

ABSTRACT

Corpus approaches to issues in second language acquisition: three studies

Randy Fred Appel, Ph.D. (ABD)

Concordia University, 2017

This dissertation demonstrates how advancements in corpus approaches to linguistic inquiry can be used to improve the methodological rigour, reliability, and general usefulness of findings in various areas of Second Language Acquisition (SLA) research. Although these studies primarily focus on improvements in areas where corpus approaches are already commonplace, this dissertation also demonstrates how corpus methods can be usefully applied to new areas. Through the use of these methods, the presented studies highlight issues learners face when attempting to gain proficiency in second language (L2) English.

Study 1 investigated the usefulness of transitional probability as a way of improving the extraction of formulaic sequences (e.g., *on the other hand*) from large scale corpora. Since current methods of identification often lead to lists of overlapping structures that lack psycholinguistic validity and pedagogical usefulness (Liu, 2012; Nekrasova, 2009; Simpson-Vlach & Ellis, 2010), this study evaluated the effectiveness of a new statistical measure in this area, transitional probability, as a way of improving the psycholinguistic status of corpus derived formulaic sequences. Using a sequence completion task, results revealed that corpus derived formulaic sequences with higher transitional probabilities were more accurately completed by first language (L1) and L2 English users, leading to the conclusion that these sequences are more likely to be stored as prefabricated units.

Study 2 used a corpus approach to investigate the relationship between L1 background and the lexical choices made by L2 English writers. Looking specifically at L2 English writers of L1 Arabic, Chinese, and French backgrounds, a corpus of 150 argumentative essays written as part of an English for Academic Purposes program at a large English-medium university in North America was used to identify production tendencies in the use of linking adverbials by each L1 group. Results revealed important L1 differences for the use of specific linking adverbials and broader functional categories.

Study 3 investigated lexical dimensions of L2 English speech associated with differences in perceived linguistic ability as judged by naïve L1 English raters. Using a corpus of transcribed speech samples from 97 L2 English users across two tasks (194 speech samples), naïve L1 English raters evaluated each sample for perceived comprehensibility and nativeness. Variables associated with factors related to dimensions of lexical density, sophistication, and diversity were targeted for potential correlations with L1 rater judgements of each construct. Results indicated important linguistic measures significantly correlated with each construct as well as task-based differences.

Acknowledgments

First and foremost, I would like to thank my family for all the kindness, love, and support I continuously receive from them. My parents have always made themselves available to me no matter how busy they are in their own lives, and without their support I would never have been able to accomplish this task. Thank you for giving me a chance to succeed in my chosen path.

I would also like to thank Dr. Kim McDonough and Dr. Pavel Trofimovich also for their incredible support throughout this process. Each of you have proven to be incredibly knowledgeable and supportive supervisors, and I greatly appreciate the many hours of work you have put in on my behalf.

Contribution of Authors

Study 1 and Study 3 of this dissertation are co-authored with my co-supervisor, Dr. Pavel Trofimovich. As the lead author, I was the primary researcher in each study and was responsible for collecting user data, conducting analyses, and writing manuscripts. Although the speech samples used in Study 3 came from a previous research project conducted by Dr. Trofimovich, rater data were collected specifically for this study. The idea for Study 1 came about as a result of my previous research involving the identification of formulaic sequences in large scale corpora, and my belief that improved methods of identification were needed. The idea for Study 3 was a result of my belief that there was a need for a better understanding of the role lexical features play in naïve rater assessments of L2 English spoken ability. Study 2 of this dissertation is single authored.

Table of Contents

List of Tables

Glossary

Formulaic Sequences: Prefabricated multi-word structures that represent single choices in the minds' of users. These sequences are believed to be stored and produced whole at the time of use and therefore associated with important processing benefits that help users produce quick and accurate discourse. Examples of formulaic sequences include *kick the bucket*, *once upon a time*, and *the fact that*.

Linking Adverbials: Lexical units that improve discourse cohesion by helping the reader interpret information that follows in light of what has already been presented. Frequently occurring at sentence boundaries, linking adverbials can also appear within sentences and are often separated from the rest of the sentence by way of punctuation markers. Examples of linking adverbials include *as a result*, *in addition*, and *consequently*.

Comprehensibility: A characteristic of L2 speech based on impressionistic judgements regarding the ease/difficulty of understanding. Comprehensibility is commonly evaluated using scalar ratings by naïve judges (Munro & Derwing, 1995; Isaacs & Trofimovich, 2012).

Chapter 1: General Introduction

Introduction

The manuscripts presented in this dissertation aim to improve our understanding of Second Language Acquisition (SLA) by looking at the lexical characteristics of first language (L1) and second language (L2) English discourse from a corpus perspective. The use of corpora as a main theme connecting these studies comes from my belief that, when used appropriately, performance data can provide important insights into L1 and L2 production patterns that may remain hidden when investigated via other means.

With the use of corpora as a uniting factor, this dissertation carries a heavy methodological focus that aims to push the field forward by demonstrating how traditionally corpus-informed areas of linguistic inquiry can be improved upon by using new statistical techniques, increased methodological rigour, and a closer adherence to established theoretical frameworks. Additionally, this dissertation demonstrates how these improvements allow corpus methods to be applied to additional areas of linguistic inquiry that have previously underused such approaches.

In order to demonstrate the value of corpus research methods and how they can be improved upon, this dissertation targeted three key areas of SLA that are well suited to this methodology: identification of formulaic sequences in large scale corpora, L1 related production tendencies in L2 English academic writing, and the quantification of lexical features correlated with assessments of L2 English speech performance. These topics are important to SLA researchers since each represents an area that has received significant attention over the last several decades and is seen as important to our general understanding of L1 and L2 acquisition.

The first issue, identification of formulaic sequences in large scale corpora, is of interest because of the growing recognition that multi-word lexical units play an important role in proficient language use and are associated with processing advantages that aid in the production of quick and accurate discourse (Conklin & Schmitt, 2008; Underwood, Schmitt, & Galpin, 2004). Unfortunately, while the importance of formulaic sequences is increasingly recognized, current methods of identifying these structures often lead to lists of incomplete, overlapping, or overly extended structures that lack psycholinguistic validity and pedagogical usefulness (Liu, 2012; Nekrasova, 2009; Simpson-Vlach & Ellis, 2010). Since we lack a proven methodology that can be used to reliably identify these sequences, we are unable to accurately measure the amount of formulaic language that appears in naturally occurring discourse, or highlight pedagogically valuable sequences that could be of use for L2 English learners. Study 1 attempted to address limitations in previous efforts to extract formulaic sequences from large-scale corpora by demonstrating the effectiveness of a new statistical measure in this area, transitional probability, which can be applied as a directional measure of word association to be used in conjunction with traditional approaches to formulaic sequence extraction. Findings from Study 1 revealed that the application of this metric can improve the psycholinguistic status, and therefore pedagogical usefulness, of formulaic sequences extracted from large scale corpora.

The second issue addressed in this dissertation concerns itself with the relationship between L1 background and L2 lexical use. More specifically, identification of L1 related production tendencies in the use of linking adverbials (e.g., *on the other hand*, *in addition*) by L2 English academic writers. While previous research on this subject has highlighted several noteworthy production tendencies, the methods most commonly employed in these studies have prevented a distinction between L1 related and universal features of L2 English from being

made. Study 2 attempted to address limitations in previous corpus research of this kind by analyzing a closely controlled corpus of 150 L2 English academic essays produced by writers from three different L1 backgrounds (Arabic, Chinese, and French). Looking specifically at the use of linking adverbials, Study 2 applied a corpus-informed approach to highlight unique production tendencies in the L2 writing of each L1 group.

The third issue, which lexical factors are indicative of differences in perceived spoken ability is an understudied area. Since written English is generally considered easier to work with than spoken English, there is a comparative lack of corpus-informed research on speech. While the influence of pronunciation, prosody, and fluency have all been explored in previous L2 based speech performance research (Derwing & Munro, 1997; Derwing, Rossiter, Munro, & Thomson, 2004; Kang, Rubin, & Pickering, 2010; Munro & Derwing, 1999), much less is known about the role lexical factors play in these judgments. To address this issue, Study 3 used a corpus of transcribed speech samples from 97 L2 English speakers, across two tasks, to analyze quantifiable differences in lexical measures that contribute to naïve L1 English raters' evaluations of comprehensibility and nativeness.

Increased knowledge in each of these areas is important since findings from these studies can influence our understanding of the language acquisition process, language teaching pedagogy, and constructs related to language assessment. Without concrete evidence to guide our views on each of these issues, we are unable to provide teachers, researchers, assessment specialists, and second language learners with the appropriate knowledge base that should act as their guide.

In the overarching review of the literature that follows I begin by providing a brief historical overview of corpus linguistics, including its loss of popularity following the

Chomskyan revolution and its subsequent reemergence in light of the advancements created by the introduction of the digital age. Following this relatively brief review, some major challenges associated with modern day corpus linguistics are discussed in relation to the three studies in this dissertation along with specific suggestions for how these challenges can be overcome.

Overarching Review of the Literature

Early Corpus Linguistics

As a label, 'corpus linguistics' is a relatively new term that only began to appear in the 1980s (McEnery, Xiao, & Tono, 2006), yet the roots of this approach can be traced back to at least the 19th century (McEnery & Wilson, 2001). Early attempts at what today would be considered a corpus approach were present in the work of field linguists who carefully organized collections of discourse recorded on individual slips of paper as a source of data for analysis (Svartvik, 1991). In the absence of digital technologies, the accumulation, organization, and storage of these discourse samples was, in itself, a time consuming process, yet without the benefit of computer aided forms of analysis, research involving these early corpora was a necessarily arduous, time consuming, and labour intensive ordeal. In fact, in order to fully analyze these large collections of text, researchers often needed to employ huge workforces to help sift through their data. For example, Kaeding (1897) is said to have employed at least 5,000 workers to help analyze his corpus of nearly 11 million words (Kennedy, 1991; McEnery & Wilson, 2001).

The labour intensive and time consuming nature of early corpus linguistic work meant that research of this type was often error prone since achieving consistency in the analysis of large collections of manually searched discourse, especially when employing a large workforce,

was near impossible. As a result, early corpus linguistics was often viewed as a pseudo-procedure (Abercrombie, 1966) that was too labour intensive, time consuming, and error prone to have a proper place in linguistic inquiry.

These technical limitations presented a powerful critique of corpus linguistics at the time and did much to discredit these methodologies, yet it was the theoretical shift caused by the Chomskyan revolution in the 1950s and 1960s, and its devaluation of performance data, that may have been the biggest contributor to the decreased popularity of corpus linguistics in most areas of linguistic inquiry after the 1950s (McEnery & Wilson, 2001; McEnery et al., 2006). With the introduction of Chomskyan linguistics and its focus on competence over performance, corpora were no longer seen as a primary source of linguistic data in many areas, and were forced to give way to more rationalist approaches to linguistic inquiry.

The rise in popularity of Chomskyan linguistics and its closely linked generative grammar theoretical framework resulted in a steep decline in corpus informed research at this time. However, the approach was never completely abandoned. Important corpus-informed research continued to take place and methodological advancements were made. The most important of these advancements, and the one that greatly contributed to the reemergence corpus linguistics as a respected research tradition, was the introduction of new technologies that allowed for the digital storage and analysis of large collections of text.

Modern Day Corpus Linguistics

Despite the many criticisms leveled at pre-digital corpus linguistics, the introduction of the computer age brought about an important maturation to this type of linguistic inquiry. With digitized texts and the accompanying automated forms of analysis that began to be developed

alongside them, corpus methods reemerged into what is now considered an important and valuable tool that can help reveal findings that lie beyond the gaze of human introspection (Stubbs, 2007). With an ever increasing reliance on computer assisted forms of analysis, the technological limitations that marred early corpus linguistic research, leading many to view it as a pseudo-procedure, were mostly left behind. Although not all of the methodological and theoretical criticisms that had been levelled at corpus linguistics were fully absolved, an increasing recognition began to take hold that, for many questions, the only valid answers that could be achieved were to be found through corpora (Mair, 1991; Stubbs, 2007). With computers now facilitating much of the data analysis, projects that had once taken days, weeks, or even years to accomplish could be completed within minutes. As a result, new possibilities began to reveal themselves and the breadth of corpus research began to grow. Since these processes were increasingly automatized, reliability of results also increased, and a noticeable boom in corpus research began to take hold beginning in the 1980s (McEnery & Wilson, 2001).

In the years following the introduction of the digital age, corpus linguists have targeted the compilation and analysis of both written and spoken discourse, yet the main focus has largely resided in the study of written language due to its stable and easily reviewable form. Corpora in these early years of modern corpus linguistics were also mostly focused on English discourse since this was the primary language of many of the pioneers in the field, and it was not until the 1990s that corpora of learner English began to appear (Granger, 1998). As a consequence of this general focus on the written word and relatively recent introduction of L2 English corpora, there is an on-going need for more corpus informed research on L2 speech.

Corpus-informed research has witnessed incredible growth since the introduction of the digital age, yet there are still many important areas that require improvement to push the field

forward. In this dissertation, the collected studies focus on three such issues that require attention as it relates to the use of corpus linguistics in SLA: the theoretical divide between corpus linguistics and SLA theories, the need for improved methodological rigour, and the expanded exploration of corpora of L2 English spoken discourse through new forms of analysis.

The Theoretical Divide

Despite the fact that corpus linguistics is now seen as a powerful research tool that can be applied to numerous areas of linguistic inquiry, the approach has largely remained a theory free endeavour (Gries, 2010; Stubbs, 2007). In fact, it has been stated that many corpus linguists would go so far as to say that theoretical linguistics and corpus linguistics stand in stark opposition and that the differences between the two lead to an irreconcilable incompatibility (Halliday, 2005). This general opposition, coupled with the occasional view that corpus linguistics is itself a theory instead of a method of inquiry (e.g., Leech, 1992; Stubbs, 1993), has led to a frequent divide between corpus linguistics and theoretical models of SLA that can have a negative impact on both areas by limiting possibilities for advancement.

While the absence of a theoretical basis can be, at times, beneficial since it allows more freedom in terms of the specific methodologies employed and interpretation of results, it can also limit the value of findings since they are unable to be placed within the larger context provided by an established theoretical framework or be used to test the assumptions present within it. In order to better guide the research being performed and provide more meaning to results, arguments have begun to emerge that suggest there is a need for a stronger theoretical basis in corpus studies so that theory and method can begin to inform each other (e.g., Gries, 2010).

In addition to helping researchers better interpret their findings, an important motivation for greater theoretical engagement is that, in the absence of a theoretical basis, corpus linguistics has come to accept a wide range of questionable research methods (Barlow, 2011). This is evident in the corpus-informed identification of recurrent word sequences. In this area of corpus research, the lexical bundle and n-gram approaches have become well-accepted standards, yet the application of these methodologies often lead to lists of overlapping structures that are difficult to functionally categorize, lack psycholinguistic validity, and are of questionable pedagogical usefulness. Unfortunately, since this research is generally performed in the absence of a theoretical basis, these issues are often ignored.

Study 1 attempts to address this issue by more closely aligning the identification of formulaic sequences in large-scale corpora with a theoretical framework that places value on multi-word utterances. In this case it is usage-based, or emergentist, models of language that are used to provide a more theoretically grounded approach to the issue. By basing the identification of formulaic sequences in this theoretical framework, the unit of focus shifts from simple recurrent word combinations to form function pairings that represent valid units of meaning (labelled as constructions in usage-based models). As a result of this theoretical underpinning, the statistical approach used to extract these sequences is necessarily modified so that the unit status of identified structures is clearer, thereby resulting in the identification of sequences that have improved psycholinguistic validity and pedagogical usefulness. In this way, Study 1 demonstrates how a greater adherence to established theoretical frameworks in SLA can be used to improve methodology and provide more value to findings achieved through corpus research.

Study 1 demonstrates the important benefits that can be gained by more closely aligning corpus methodology with theories of SLA. However, improvements in other areas of corpus

research are also needed to help push corpus linguistics forward. While the shift from physical to digital storage and analysis has helped increase methodological rigour by improving reliability and validity, rigour is an ongoing area of concern. In the following section, issues of methodological rigour are discussed in reference to the, at times, confusing area of corpus informed research on L1 related differences in L2 English writing.

Methodological Rigour

Methodological rigour is an area of concern in any form of research. Without sufficient rigour, the validity of results is lessened and the applicability of findings decreases. Although methodological rigour can be considered a crucial requirement in all research, and is therefore addressed by each of the three studies in this dissertation, it is most effectively expressed in reference to Study 2 which focuses on the identification of unique production tendencies in the use of linking adverbials by L2 English academic writers of three L1 backgrounds.

Due to the important role linking adverbials play in discourse cohesion, as well as the difficulty L2 learners frequently face in regards to appropriate and effective use, a growing body of research has emerged that aims to determine how these items are used by L2 English writers of various linguistic backgrounds. In previous studies on this subject, researchers have primarily relied on native/non-native contrasts to highlight patterns of overuse, underuse, and misuse related to specific groups of L2 English users (e.g., Carrio-Pastor, 2013; Chen 2006; Lei, 2012). While these studies have suggested several potential L1 related production tendencies, and resulted in occasional claims for L1 background as the underlying cause of identified differences, it remains to be determined which production tendencies are in fact associated with particular L1 groups, which can be more universally associated with L2 English learners of all linguistic

backgrounds, and which may simply have been caused by methodological limitations present in these studies.

In terms of methodological limitations, lack of corpus comparability has repeatedly appeared as an important area of concern in this body of research (e.g., Altenberg & Tapper, 1998; Chen, 2006; Granger & Tyson, 1996; Milton & Tsang, 1993). While comparability of corpora should be considered a crucial prerequisite in writer contrasts, studies of linking adverbials in L2 English academic writing have largely failed to control for important differences in target language proficiency, writing conditions, writing type, and sample length – all of which may impact linking adverbial production. As a result, it is unclear whether production tendencies identified in these studies are a result of the L1, differences in target language proficiency, or any other uncontrolled factors. For example, Milton and Tsang (1993) compared a collection of L2 English academic essays to the Brown and London Oslo/Bergen (LOB) corpora. Since the Brown and LOB are both composed of general English, and largely lack the inclusion of any significant amount of academic discourse, their use represents a genre incongruence that may have led to a misidentification of production tendencies.

Additionally, as it relates to the goal of distinguishing between unique L1 related production tendencies and universal features of L2 discourse, the prevailing focus on native/non-native contrasts must also be addressed. Since studies implementing this approach generally aim to compare L1 users with L2 writers of a single linguistic background (e.g., Bolton et al., 2002; Carrio-Pastor, 2013; Chen, 2006; Lei, 2012), they lack the ability to distinguish between production tendencies attributable to all L2 learners and those related to a specific L1 group. Put simply, by targeting L2 users of a single L1, it is impossible to ensure that identified production tendencies are in fact unique to the particular group of L2 writers being targeted.

Study 2 attempts to overcome methodological limitations in previous research of this kind by using a specially designed corpus that controls for differences in target language proficiency, writing conditions, and essay type to identify examples of intra-group homogeneity and inter-group heterogeneity in the writing produced by L2 English learners from three different L1 backgrounds (Chinese, French, and Arabic). Thus, Study 2 attempts to use increased methodological rigour to more accurately distinguish between unique production tendencies associated with particular L1 groups and universal features of linking adverbial production that are common to all L2 English writers.

The preceding discussion has concentrated on the importance of methodological rigour in corpus informed studies of L1 related production tendencies, and therefore has been made in reference to Study 2. However, each of the studies in this dissertation considers methodological rigour to be of major importance and attempts to address this issue in its own way. As such, this dissertation demonstrates how improved methods can add value to corpus informed approaches in SLA. While Studies 1 and 2 focus on methodological improvements in well-established areas, Study 3 demonstrates how technological advancements allow corpus approaches to be applied in new ways to alternative areas of applied linguistics research.

New Directions

Although corpus approaches are widely used in applied linguistics, in general, corpus linguists have tended to favour analyses of written discourse due to the relative ease associated with the collection of data in this register. As a result, corpus studies of oral discourse are comparatively rare and more corpus-informed, speech-based research is needed. Coupled with the fact that corpora of learner English only began to appear in the 1990s (Granger, 1998), and

that the analysis of L2 English has also carried a heavy focus on writing, there is an ongoing need for more corpus-informed studies targeting L2 English speech.

Although the analysis of L2 oral discourse can take on many forms, recent advancements in automated assessment and other forms of computer assisted analyses have helped create new ways of looking at performance data that were previously either prohibitively time consuming or too labour intensive to be used on a large scale. As an added benefit, the use of these software programs has also led to greater stability and objectivity in corpus-informed studies by largely eliminating the need for manual analysis that can lead to errors and personal bias. In Study 3, two computer aided forms of lexical analysis, Coh-metrix 3.0 (Grasesser, McNamara, Louwerse, & Cai, 2004) and VocabProfile (Cobb, 2016), are used to provide a quantifiable understanding of the lexical features associated with perceived differences in L2 English oral ability.

Coh-metrix, freely available in online form, allows researchers to use sophisticated software to better understand a wide range of lexical characteristics appearing in individual discourse samples. Through the use of this online tool, researchers can quickly calculate data on 11 broad categories, divided into 108 specific measures, to quantify a wide variety of discourse features. VocabProfile, also available free in online form, allows researchers to compare user-supplied discourse samples with various preestablished word lists to provide an indication of the level of lexical sophistication present in each sample. In Study 3, Coh-metrix and VocabProfile are used to analyze a corpus of speech samples composed by 97 L2 English users of varying linguistic abilities. By transcribing this corpus of L2 English speech, individual samples can be submitted to these online systems and quickly analyzed for a range of lexical features.

With transcription of samples a prerequisite step in the use of these online tools, they carry the added benefit of allowing for analysis of speech from a purely lexical perspective. This

is because other confounding factors that typically contribute to assessments of L2 English speech (e.g., segmentals, prosody, fluency) are eliminated during the transcription process. Thus, by using transcribed samples, Study 3 is able to focus exclusively on how lexical features in L2 English speech contribute to assessments of linguistic ability. Since this is an area that has only recently begun to be investigated in this manner, Study 3 is able to explore a growing area of research and demonstrate the benefits of this new approach to discourse analysis.

Tying it All Together

Each of the manuscripts in this dissertation carries its own unique methods and perspectives, yet they are all tied together by a common belief that sees corpus data is a valuable starting point in the analysis of L1 and L2 discourse. Although the introduction of the digital age in corpus research has created an ever increasing dependence on computer aided forms of analysis, and led several scholars to view automated measures as an integral part of corpus research, this dissertation takes a broad view of corpus linguistics that includes a wide range of manual and automated techniques. Therefore, as used in this dissertation, the term *corpus linguistics* refers to any approach that makes primary use of performance data to answer relevant research questions. The decision of which specific tool and method to apply is dependent on the nature of the corpora being investigated and the particular research questions to be answered (Huston, 2002).

In studies 1 and 3, computer aided forms of analysis are seen as necessary, as their application results in improved consistency and accuracy when extracting measures for a wide range of linguistic features (Study 3) and from a large data set (Study 1). Without computer aided techniques to automate extraction, analyses in these studies would be overly time consuming and error prone, thereby degrading reliability and validity of the research. In contrast,

Study 2, which uses a manual approach to analyze a small and highly specialized corpus of L2 discourse, benefits from the added level of discretion and human judgment made available by manual analyses, as the targeted lexical category (linking adverbials) is frequently misidentified by automatic measures. Thus, the use of manual analyses in Study 2 is a necessary step that helps to control for inaccuracies that can result from more automated methods of extraction. By targeting both manual and automatic forms of analysis, this dissertation aims to highlight the wide range of approaches available to corpus linguists.

Despite a common methodological focus on corpus approaches to linguistic inquiry, this dissertation does not simply attempt to replicate previous research, but instead make incremental changes that, hopefully, improve the usefulness of the findings and the applicability of the results arrived at in each area of study.

Introduction to Study 1

Study 1 of this dissertation investigated the topic of corpus-informed formulaic sequence extraction. The goal of Study 1 was to improve methods of identifying these sequences in large scale corpora so that more theoretically valid and psycholinguistically real structures could be identified. To achieve this goal, Study 1 used usage-based/emergentist models of language as an overarching theoretical framework to guide the methodological approach of the study. By aligning the construct, formulaic sequences, with an appropriate theoretical framework, usage-based/emergentist models, Study 1 attempts to demonstrate the methodological benefits that can be gained by more closely aligning corpus linguistic methods with accepted theoretical frameworks in L1 and L2 acquisition.

Chapter 2: Study 1

Transitional probability predicts native and non-native use of formulaic sequences

Published in International Journal of Applied Linguistics

By Randy Appel & Pavel Trofimovich

Abstract

Formulaic sequences (FSs), or prefabricated multi-word structures (e.g., on the other hand), are often difficult to identify objectively, and current corpus-driven methods yield structurally incomplete, overlapping, or overly extended structures of questionable psychological validity and pedagogical usefulness. To address these limitations, this study evaluated transitional probability as a potential metric to improve the identification of FSs by presenting 100 four-word sequences from the British National Corpus, varying in transitional probabilities between words, to native and non-native speakers of English (N = 293) in a sequence completion task (e.g., for the sake __). Results revealed that the application of transitional probability reduces many of the problems associated with current approaches to FS identification and can produce lists of FSs that are more functionally salient and psychologically valid.

Introduction

The study of formulaic language has been approached from a variety of perspectives following Firth's (1935) assertion that words tend to co-occur in particular patterns. It is now a well-known fact that spoken and written language includes predictable multiword units carrying specific meanings and that these units form a core component of natural language use (Schmitt, 2010). According to usage-based approaches to language learning (e.g., Barlow & Kemmer, 2000), recurrent units of meaning are represented by usage events or constructions, which refer to pairings of form and meaning (e.g., *I don't know*, *kick the bucket*), and it is the exposure, categorization, and subsequent probability assessments of these events over time that lead to language learning. As exposure to usage events increases, through input and output, probabilities related to the acceptability of utterances are accumulated and eventually used to interpret and produce discourse. Over time, frequent sequences can come to be stored as prefabricated units that are pulled from memory fully formed at the time of use. These structures, referred to as formulaic sequences (FSs), make up a large portion of discourse (Erman & Warren, 2000; Schmitt & Carter 2004), contribute to fluent, nativelike speech (Kuiper, 1996; Pawley & Syder, 1983;), characterize proficient language use (Bamberg, 1983; McCauley, 1985), and confer processing advantages in comprehension and production (Peters, 1983; Tremblay, Derwing, Libben, & Westbury, 2011).

With an increased recognition of FSs, research targeting their identification has also grown. However, since formulaic language comes in many forms with varying lengths, degrees of fixedness, levels of grammatical acceptability, and semantic opaqueness, there is no single definition of formulaicity. As a result, researchers have developed different criteria, such as frequency statistics and association indexes, and created a variety of terms (e.g., chunks,

amalgams, prefabricated routines) to label the object of study. Thus, terminology is difficult to reconcile across studies, largely because different terms and criteria are applied to the same general concept. Among the most easily recognizable FSs are idioms, such as *a stitch in time saves nine* and *kick the bucket*, due to their semantic opaqueness and non-compositionality (i.e., the meaning of these structures cannot be inferred from the combination of their constituent components). However, other frequent fixed-form FSs, such as *on the other hand*, *by the way*, and *the fact that*, are difficult to label consistently because they serve various discourse functions, are often more semantically clear and compositional, and can be identified according to various criteria.

The goal of the current study was to evaluate a novel corpus-derived criterion for improving the identification of fixed-form FSs, such as *on the other hand* and *by the way*, with the aim of more accurately identifying psychologically valid sequences that are more functionally salient, theoretically grounded and teacher friendly. As discussed below, current corpus-driven methods yield structurally incomplete, overlapping, or overly extended structures of questionable psycholinguistic validity, functional salience, and pedagogical usefulness (Liu, 2012; Nekrasova, 2009; Simpson-Vlach & Ellis, 2010). Therefore, it is important to develop and test new methods of objectively identifying these structures. Doing so will enable us to understand recurrent word patterns present in various corpora that may otherwise have gone unnoticed. Simply relying on personal reflection or subjective judgements of formulaicity is insufficient in this regard, since many recurrent word patterns only begin to emerge through analyses of large amounts of text. Developing accurate methods of identifying FSs is also important because FSs can facilitate language acquisition. Indeed, FSs represent a large portion of native-speaker discourse (Erman & Warren, 2000) and focused instruction targeting FSs may

help learners incorporate them into their linguistic repertoires (Boers, Eyckmans, Kappel, Stengers, & Demecheleer, 2006). Thus, to identify fixed form FSs and evaluate their use, this study targeted the statistical measure of transitional probability, a previously unused metric in this field. The assumption was that transitional probability, which assesses word association strength to indicate utterance boundaries, should lead to more accurate FS identification (i.e., with fewer incomplete, overlapping, and overly extended structures) and better predict FS use in both native and non-native users, compared to other criteria, such as frequency and mutual information statistics.

Corpus-driven research into identification of FSs

Corpus-driven research, with its frequent focus on large repositories of spoken or written language, has helped to remove much of the subjectivity associated with native-speaker judgements, used previously to define FSs (Conrad, 2000). This research largely targets two separate yet associated approaches, namely, *n-grams*, identified according to frequency of occurrence, and *lexical bundles*, identified using frequency and range. However, both methods assume that frequency dictates importance of the structure, and this assumption is illustrated through research on lexical bundles, an increasingly popular method of identifying FSs in various genres and registers.

Defined as "the most frequently recurring sequences of words" (Biber & Barbieri, 2007, p. 264), lexical bundles refer to a subset of formulaic language derived through frequency and range requirements. Identification of lexical bundles involves the analysis of digitized text, sourced from written or oral discourse, through concordancing software (e.g., Wordsmith Tools, Collocate) which scans the text for repeated structures. With the size of the 'window' (sequence length) usually set at four words (Biber & Barbieri, 2007; Cortes, 2006; Hyland, 2008), the

software scans the text beginning with the first word in the corpus. As the window moves forward, it takes four-word 'pictures' of the corpus in a progressive manner, such that words 1–4 represent the first sequence followed by words 2–5, 3–6, 4–7 and so on. Each sequence is recorded, and repetitions are tallied to create a list of recurrent sequences. With minimum frequency threshold often set at 20–40 occurrences per million words, the software can produce lists of sequences meeting this criterion. Generated sequences are subsequently reviewed to ensure adherence to a minimum range requirement, typically set at 3–5 texts (e.g., Biber & Barbieri, 2007) or 10% of texts used in the corpus (e.g., Hyland, 2008), to eliminate frequent sequences confined to limited texts produced by individual speakers or writers.

The lexical bundle methodology also carries limitations that result in overlapping, overly extended, and structurally or semantically incomplete sequences lacking psychological salience (Liu, 2012; Nekrasova, 2009; Simpson-Vlach & Ellis, 2010). Because this approach relies almost exclusively on the frequency criterion, it often identifies repeated structures that carry little actual meaning. For example, two frequent FSs that appear together in discourse (e.g., *due to* and *the fact that*) will often be presented by the concordancing software as multiple four-word entries with no unitary semantic status (i.e., *due to the fact*, *to the fact that*, *the fact that the*). Because sequence length is determined *a priori* by the researcher, the process generates lists which misrepresent complete structural units, such as sequences that cross syntactic boundaries, often with a determiner (*a/the*) in terminal position (e.g., *the fact that the*, *the nature of the*). In these cases, the determiner likely belongs to a separate unit of meaning and should not be included with the structure. The often incomplete semantic and structural status of lexical bundles also contributes to difficulties with the assignment of functional roles. Cortes (2004), for instance, classifies *the fact that the* and *the nature of the* as discourse organizing bundles. However, Biber,

Conrad, and Cortes (2004) view them as serving stance and referential purposes, likely as a result of the utterance-final *the* referring to a new unit of meaning.

Additional problems with lexical bundles relate to sequence length. With lexical searches often limited to four-word units, the assumption is that all FSs include four words. For instance, it is often argued that four-word bundles "offer a clearer range of structures and functions than 3-word bundles" (Hyland 2008, p. 8) and that four-word sequences subsume three-word units (Cortes 2004). However, as shown above, the functions of many four-word bundles do not seem as clear, and the fact that four-word structures contain three-word units is a drawback, not an advantage. Finally, problems with the identification of lexical bundles also question their psychological status as prefabricated units. Since lexical bundles are identified based on frequency, they often cross syntactic and semantic boundaries and lack clear meanings (i.e., *and one of the*, *going to be the*), which would not normally be associated with usage events in the mind of a language user. This also raises questions about the utility of lexical bundles as pedagogical tools since it would be hard for learners to use FSs which lack clear functional roles.

To sidestep some of the limitations of the lexical bundle approach, researchers recently proposed another statistical measure – termed mutual information index – as a supplement to existing frequency and range criteria (Simpson-Vlach & Ellis, 2010). Mutual information refers to the ratio of the observed frequency of an entire n-word sequence in a corpus (e.g., *cup of tea*) relative to the expected frequency of that same sequence occurring by chance alone. Mutual information encompasses probability values for each constituent word in the target structure, with the total probability being a product of all individual word probabilities. The assumption underlying the mutual information statistic is that frequency alone often fails to identify important word associations and that mutual information, with its focus on associations between

words, can yield more functionally salient and structurally complete FSs. For example, Simpson-Vlach and Ellis (2010) successfully used mutual information to improve functional salience and significantly reduce the number of identified structures with a determiner in terminal position.

Unfortunately, mutual information does not take into consideration word order, so it can only be used to calculate co-occurrence of an entire sequence, not sequential probability. For instance, the mutual information values for *cup of tea*, *tea of cup*, *cup tea of*, and *tea cup of* are all identical, illustrating the main limitation of this measure, namely, its insensitivity to sequential probabilities of occurrence between elements. Additionally, as identified by Biber (2009), mutual information tends to disfavour sequences that contain high-frequency function words, such as *the*, *of*, *a*, with the consequence that mutual information values become disproportionately reduced for target sequences with highly-frequent words. Therefore, although mutual information has been used to improve some aspects of corpus-driven identification of FSs, more effective approaches are still needed.

Transitional probability

With the overall goal of developing a more effective approach to corpus driven identification of FSs, this study targeted the measure of transitional probability which has been used in psycholinguistic research on word segmentation and statistical learning (e.g., Aslin & Newport, 2012; Mirman, Estes, & Magnuson, 2010). Transitional probability is a measure of co-occurrence of segments, syllables, or words in a sequence, estimating the likelihood of a particular element being followed by another. Forward transitional probability, the likelihood of X being followed by Y, establishes the frequency of XY relative to all occurrences of the initial element in the sequence. Similarly, backward transitional probability, the likelihood of X preceding Y, denotes the frequency of XY relative to all instances of the final element in this

sequence. In word segmentation research, high transitional probability between syllables suggests that syllables co-occur and likely represent a word-like unit, while low transitional probability between syllables implies word boundaries. Both children and adults can compute such statistics after about two minutes of exposure to a novel sequence, demonstrating above-chance segmentation performance (Aslin & Newport, 2012). This study extends this notion from research on speech segmentation to research on formulaic language, to more accurately isolate utterance boundaries and ultimately achieve more precise FS identification.

While similar to mutual information in that it can be used to compare probabilities of occurrence relative to the frequencies of individual elements, transitional probability holds the advantage of taking into account sequence order when making these calculations and can be used to measure the strength of association at different points in a structure. This is an important benefit, suggesting that transitional probability can be combined with standard lexical bundle and n-gram methodologies to better understand the boundaries of FSs and thus more accurately identify which sequences function as fixed-meaning units, helping reduce the incidence of overlapping, incomplete, and overly extended structures identified as FSs.

The current study

In light of the numerous limitations associated with the traditional lexical bundle approach and the more recent introduction and use of mutual information statistics, this study evaluated transitional probability as a new method of improving the identification of fixed-form FSs in large corpora. This study, whose goal was to test the effectiveness of transitional probability as a metric of FS status with both native English users and second language (L2) learners (i.e. individuals for whom psychologically valid and pedagogically useful FSs will arguably be most relevant), was guided by two closely related research questions:

1) Can the application of forward and backward transitional probability be used to improve the identification of functionally salient and psychologically valid formulaic sequences that are better suited for pedagogical purposes?
2) To what extent does transitional probability, compared to the traditional frequency of occurrence and mutual information statistics, predict native English users' and L2 learners' completion of highly frequent four-word sequences?

The answer to these questions will reveal a greater understanding of how FSs function within discourse and help improve our ability to objectively identify these structures in large corpora. Although a limited range of FSs have begun to appear in some materials aimed at L2 English users, English for Academic Purposes (EAP) materials often underrepresent this aspect of language and do not provide L2 users with the kinds of FSs they are likely to face in their academic careers (Wood & Appel, 2014). It is therefore important to investigate methods of improving the identification of pedagogically useful FSs that can be incorporated into language teaching materials aimed at a variety of L2 English users' needs.

To evaluate the effectiveness of transitional probability in the identification of FSs, 100 frequently occurring four-word sequences were extracted from the British National Corpus, with the idea that some represented 'fixed' four-word FSs but others encompassed partial or incomplete structures that are misidentified by applying the traditional lexical bundle approach. These sequences were presented to 138 native English speakers and 155 L2 English learners in a sequence completion task, with the fourth word deleted, to examine the rate and variability in sequence completion. Using both native and non-native English users was crucial in testing the utility of transitional probability for research and teaching purposes, on the assumption that the FSs identified through transitional probability should ultimately be useful for L2 learners, who

(as argued above) might benefit from psychologically and pedagogically valid FSs in language learning. These scores were tested against the transitional probability, mutual information, and traditional frequency of occurrence statistics to determine which measure best predicted language users' performance.

Method

Participants

Participants included 138 native English speakers and 155 L2 English learners. The native speakers (81 female, 57 male), who were on average 37.9 years old (20–75), resided in the United States because recruiting native English speakers with no multilingual experience and little knowledge of additional languages was problematic in Montreal, Canada, a bilingual French-English city with a large multilingual population. Therefore, native speakers were recruited through online media and tested in a timed, on-line setting. These participants all identified English as their native language, with several reporting basic knowledge of Spanish (7), French (3), Vietnamese (2), American Sign Language (2), Indonesian, Portuguese, Hungarian, and Korean (one each). They had all completed at least a high school education, with 74 and 9 holding further undergraduate or graduate degrees, respectively. They self-reported a mean of 99% of daily language use being in English (50–100%), with 96% of all interactions (40–100%) occurring with other native English speakers, using 0–100% scales (0% = *never*, 100% = *all the time*). Native speakers were allowed to complete either one or both of the two non-overlapping versions of the target materials (see below), with 122 participants completing Version A, 123 participants completing Version B, and a total of 104 responding to both. In total, each of the 100 target items was responded to by a minimum of 122 of native speakers.

The L2 English learners included 155 undergraduate students in an English for Academic Purposes (EAP) program at an English-medium university in Montreal. These participants were sourced from six intact classes in the upper-intermediate level of this program. As university students, all speakers had taken either TOEFL iBT or IELTS tests, demonstrating at minimum a total score of 85 for TOEFL iBT or 6.5 for IELTS, which was considered sufficient for them to pursue academic degrees. By including only those students from a similar course level, we were able to evaluate a relatively homogenous group of L2 users, in terms of proficiency, focusing on their ability to accurately complete the target sequences. L2 learners (87 female, 68 male) came from a variety of language backgrounds, including Chinese (89), Arabic (26), French (14), Farsi (7), Korean (5), Spanish (3), Greek (3), Turkish (2), Polish, Italian, Azeri, Marathi, Bulgarian, and Hausa (one each). While the learners came from a variety of language backgrounds, linguistic background was not a focus of this study. In fact, variability in learners' linguistic backgrounds was seen as a strength, allowing us to examine how a broad range of learners from multiple linguistic backgrounds respond to L2 FSs. L2 learners, who were on average 21.5 years old (18–37), reported a mean of 8.5 years (1–20) of prior English study. Using the same scales, they estimated their English ability at a mean of 6.9 (3–9) in speaking, 7.6 (4–9) in listening, 7.1 (3–9) in reading, and 6.5 (3–9) in writing, and reported communicating with native English speakers on average 52% of the time daily (10–100%). L2 learners, who were tested using identical but paper based materials in a timed setting, were also randomly assigned to Versions A or B of the materials, with 77 L2 learners completing Version A and 78 completing Version B. Thus, each of the 100 target items was responded to by at least 77 learners.

Materials

The target materials included 100 frequently occurring four-word sequences from the British National Corpus (BNC). The BNC was chosen since it represents a large collection of English that can easily be searched via publicly available on-line tools. Because frequency of occurrence has been used in previous corpus-driven research as a main index in FS identification, the initial step in selecting the sequences was to generate a list of 300 most frequent four-word structures in the BNC using the on-line interface at phrasesinenglish.org (Fletcher, 2011). From this list, 100 target structures were chosen through semi-random sampling without replacement, with the criterion that they represented a wide range of values for four metrics: (a) forward transitional probability; (b) backward transitional probability; (c) mutual information; and (d) frequency. Sequences deemed to be overly context specific or unique to British English were removed from the test materials. Since target sequences were taken from a large collection of British English, and the participants were speakers/learners of North American English, this was a necessary step to limit the potential impact of differences in language variety between North American English and British English. Table 1 summarizes FS statistics for 100 target sequences.

Table 1

Descriptive Statistics for 100 Target Sequences

Corpus-derived measure	*M*	*Median*	*Min*	*Max*
Frequency	1131.06	867.00	657.00	6875.00
Mutual information	13.21	13.04	5.17	23.61

Forward transitional probability	.50	.38	.03	1.00
Backward transitional probability	.67	.80	.05	1.00

The target structures included various options in terminal position, such as *of, a*, *which*, *hand*, *possible* (which were to be completed by participants) to prevent participants from completing sequences based on highly-frequent responses. The 100 target structures were subsequently organized in two randomly-sampled non-overlapping tests (Versions A & B), composed of 50 target sequences and 10 extra fillers drawn from the original list of 300 sequences, for a total of 60 sequences. In each test, all items were listed in a random order with the first three elements in each sequence printed intact and the last element replaced with a blank (e.g., *turned out to*___, *as soon as*___, *a great deal* ___). Sample target materials appear in Appendix A, and the complete list can be obtained by e-mailing the corresponding author.

Procedure

Procedures for each participant group varied slightly due to the medium in which the task was administered. The native speakers, tested in the on-line setting, first read a brief project description and digitally signed the consent form by electing to proceed to the online test materials. They were given a maximum of 20 minutes to complete a language background questionnaire followed by the sequence completion task. Before proceeding to the task, they were instructed, in writing, that they would see three-words, followed by a space to write any one word that comes to mind to complete each phrase, and that such words can be long words (e.g., *picture*, *day*, *go*, *thinking*, *fast*, *break*) or short grammatical words (e.g., *at*, *to*, *from*, *the*, *he*, *they*, *she*, *a*). They were then given examples showing possible completions for four

unrelated sequences (e.g., *for a long time*, *in the absence of*). The participants then proceeded to type in the missing words for each sequence, working at their own pace, with all completing the test within the allotted period. After finishing one test version (A or B, which the participants accessed first with an approximately equal frequency), they were invited to complete the other version (B or A, respectively), which 104 of the native speakers in fact elected to do, with no restrictions on how soon the second test version could be completed.

The L2 learners, tested as part of several intact groups of 20–25 students during their ESL classes, followed a similar procedure, except that all responses were collected on paper rather than on-line. The learners first read and signed a consent form, then completed the same questionnaire, followed by one of the test versions, for which they were also given a maximum of 20 minutes. The learners received the same instructions, both orally from the experimenter and in writing in their booklets, and were given the same four examples of sequence completions before starting the task. All learners completed the task within the allotted time. Because the learners were tested after all native speaker data had been collected and because in-person testing, compared to the on-line medium, allowed for more flexibility in study design, the ESL classes were randomly assigned to one of the two test materials (Version A or B), which resulted in nearly equal numbers of learners responding to each test version. Each learner completed only a single version of the test to reduce learner fatigue effects, eliminate missed or incomplete data points, and allow for testing to be completed within a reasonable time in a language class.

Corpus-based statistics

Four statistics were derived for each target sequence. First, frequency figures, which served as the basis for the computation of all statistics, were sourced from the BNC using the on-line interface at phrasesinenglish.org (Fletcher, 2011). Frequency was recorded as the listed

frequency count for each target structure. Second, mutual information (MI) figures were computed using the formula MI = log2 (observed frequency/expected frequency). While observed frequency corresponded to the listed frequency count in the BNC, expected frequency required additional calculations. As a first step, probability figures for each of the component words were calculated by taking the frequency of each word and dividing it by the total number of words in the corpus (Schmitt, 2010). These figures were then multiplied to achieve the overall probability of all the component words co-occurring. The resulting probability score for the entire sequence was then multiplied by the total number of words in the corpus to derive an expected frequency count for the entire sequence so that the final ratio could be computed. To illustrate this computation with a simple collocation *prime minister* (total frequency = 9,457; *prime* frequency = 11,959; *minister* frequency = 23,401), the probability of *prime* occurring in the corpus is 0.000123 (11,959/96,986,707) while the probability of *minister* is 0.000241 (23,401/96,986,707). Multiplying the two figures yields a general probability of 2.97512E-08, and multiplying this figure by the total number of words in the corpus yields the expected frequency of occurrence (2.89 times), with the resulting MI index of 11.68 (log2 [9,457/2.89]). This index is high, suggesting that the collocation appears much more frequently than would be expected by chance alone and that the component words are strongly associated.

Finally, two measures of transitional probability were computed. Backward transitional probability (BTP), the probability of the final three words in a structure appearing with the first word, was calculated using the formula BTP(X|Y) = frequency(XY)/frequency(Y), where the numerator denotes the frequency of the entire four-word sequence, and the denominator represents the frequency of the final three words in the same sequence. For example, *the fact that the* appears 2,500 times in the BNC, with *fact that the* having a frequency of 2,666. Using the

above formula, BTP equals 0.94(2,500/ 2,666), which is high, suggesting that *fact that the* is likely to be preceded by *the*. Similar to BTP, forward transitional probability (FTP), or the probability of the first three words in a sequence being followed by the fourth word, was calculated using the formula FTP(Y|X) = frequency(XY)/frequency(X), where the numerator denotes the frequency of the entire four-word sequence, and the denominator represents the frequency of the first three words. Using the same example, with *the fact that* appearing 12,987 times, the FTP statistic for *the fact that the* is 0.19 (2,500/12,987), which is low, suggesting that the final *the* is only loosely associated with the first three words in the structure. Since BTP and FTP provide distinct measures of association at different points in structures, both were used to investigate their impact on language users' ability to complete the target sequences.

Analysis

Two dependent variables were used to evaluate the effectiveness of the four metrics to predict language users' ability to complete the target sequences with the intended word. The first variable, proportion of accurate completions, was a measure of how closely participants' completions matched the terminal word in each target structure in the BNC. In the case of minor spelling errors (i.e., *thee* vs. *the*), the answer was changed to the appropriate form and counted as correct. Other answers which mismatched BNC data were counted as wrong. For each target sequence, proportion of accurate completion was derived by dividing the total number of correct answers by the total number of responses available for that sequence (e.g., 0.14 would correspond to 21 correct completions from 155 participants), separately for native speakers and L2 learners. The second dependent variable, range, was a measure of variability in participants' responses. For each target sequence, range was defined as the total number of unique responses given by at least three participants (e.g., 7 would correspond to 7 different completions to a given

sequence in three or more participants' responses), computed separately for the two participant groups.[1] Because frequency-based metrics of formulaic status, computed for each target sequence, were the focus of this study, all statistical comparisons were based on item-based statistics.

Results

Preliminary Analyses

Before examining the relationship between four corpus-derived metrics (frequency of occurrence, MI, BTP, FTP) and participants' responses in the sequence completion task, two preliminary analyses were carried out. The first focused on the overall performance of the two groups, which is summarized in Table 2. Compared to L2 learners, native speakers were significantly more likely to complete the sequences with the target word, $t(99) = 5.97$, $p < .0001$, d (effect size) = .63, with a broader range of responses provided, $t(99) = 5.10$, $p < .0001$, d = 1.03. This confirmed the expected difference in linguistic experience between the two groups, namely, that native speakers were overall more accurate, with a wider range of possible response options available to them, in providing sequence completions consistent with corpus data.

Table 2

Descriptive Statistics for Participants' Performance in the Sequence Completion Task

Dependent variable	Native speakers				L2 learners			
	M	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*
Accurate completion rate	.47	.33	.00	.99	.36	.33	.00	1.00
Range	5.70	3.51	1.00	14.00	4.33	2.32	1.00	10.00

The second analysis examined possible relationships among the four metrics by computing Pearson correlation coefficients between them for the 100 target sequences (see Table 3). With one exception, all measures were correlated with each other, suggesting that they captured a dimension common to all sequences. However, the strength of these relationships was moderate at best. For instance, frequency of occurrence and transitional probability shared only 4-9% of common variance, which confirmed that the two metrics were largely independent of each other. The overlap in shared variance between MI and transitional probability statistics was greater, yet far from perfect (8-25%), implying that the two metrics captured somewhat different dimensions characterizing the target sequences. Unlike MI, transitional probability is sensitive to the relative order of elements within a sequence, which likely reflected some unique variance (75-92%) in each measure.

Table 3

Pearson Correlations between Corpus-Derived FS Metrics

FS metrics	Frequency	MI	FTP	BTP
Frequency	—			
MI	.12	—		
FTP	.30**	.50**	—	
BTP	.21*	.29**	.52**	—

Note. MI = mutual information, FTP = forward transitional probability, BTP = backward transitional probability. *$p < .05$, **$p < .01$ (two-tailed).

Corpus-Derived Metrics as Predictors of Sequence Completions

To determine the relative contribution of the four corpus-derived metrics to native

speakers' performance in the sequence completion task, two separate stepwise multiple regression analyses were carried out, with accurate completion rate and range as criterion variables. In these analyses, summarized in Table 4, the four metrics (frequency of occurrence, MI, BTP, FTP) were used as predictor variables. For accurate completion rate, the regression model yielded a single significant predictor, the FTP statistic, accounting for 65% of the total variance. For range, the final model accounted for 64% of the total variance, with all of the variance again linked to FTP. In essence, for native speakers, FTP (i.e., the likelihood that the first three words in each sequence are followed by the final word) appeared to singly predict greater proportion of target sequence completions and lower variability associated with these completions.

Table 4

Results of Multiple Regression Analyses Using the Three Corpus-Derived Metrics as Predictors of Proportion of Accurate Completions and Range Scores for Native Speakers

Predicted variable	Predictor	*Adjusted* R^2	R^2 *change*	β	*t*	*p*
Accurate completion rate	FTP	.65	.65	.81	13.72	.0001
Range	FTP	.64	.64	-.80	-13.38	.0001

Note. FTP = forward transitional probability.

A comparable set of regression analyses was computed next, with L2 learners' accurate completion rate and range used as criterion variables. These analyses, summarized in Table 5, yielded similar findings. For completion rate, the model revealed a single predictor, FTP, accounting for 50% of the variance in learners' sequence completions. For range, the final model

explained a total of 46% of the variance, with 42% linked to FTP and a further 4% associated with the MI statistic. Although the amount of total variance explained by the L2 models was lower compared to native speaker models, the pattern of findings was similar. For L2 learners, as was the case with native speakers, FTP nearly exclusively predicted greater proportion of target sequence completions and smaller range of responses associated with these completions.

Table 5

Results of Multiple Regression Analyses Using the Three Corpus-Derived Metrics as Predictors of Proportion of Completion and Range Scores for L2 Learners

Predicted variable	Predictor	*Adjusted* R^2	R^2 *change*	β	*t*	*p*
Completion rate	FTP	.50	.50	.71	9.86	.0001
Range	FTP	.42	.42	-.65	-8.45	.0001
	MI	.46	.04	-.25	-2.95	.004

Note. FTP = forward transitional probability, MI = mutual information.

Response Consistency

The final analysis focused on participants' response agreement in completing the target sequences to determine if a given participant's behavior was consistent with the other participants in the same group. Based on the results of previous analyses, the assumption was that target sequences featuring higher FTP should elicit greater internal consistency within each participant group. For this analysis, both the overall percent of agreement as well as Cohen's kappa (κ) as an index of interrater reliability appropriate for nominal data were computed

separately for each group. Percent agreement and Cohen's kappa were calculated for each pair of participants, then averaged to yield a single value (Light, 1971). The results of these analyses, shown in Table 6, suggested that response consistency was indeed strongly associated with FTP. According to Landis and Koch's (1977) guidelines for interpreting kappa values, native speakers showed "moderate" agreement ($\kappa = .60$) for the 33 target sequences featuring high FTP values (above .70), while their agreement for the 67 sequences of low FTP (below .70) was "slight" at best ($\kappa = .19$). This relationship between response consistency and transitional probability was even more pronounced for the 11 sequences from the top and bottom of the transitional probability range, where native speakers demonstrated substantial agreement ($\kappa = .72$) and virtually random response patterns ($\kappa = .09$), respectively. As shown in Table 6, L2 learners' response consistency patterned in a similar manner, with the exception that L2 learners, predictably, demonstrated overall lower consistency and that they were less sensitive, in their response agreement, to sequences from the top range of transitional probability.

Table 6

Participant Consistency Indexes for Completion of Target Sequences of High and Low Forward Transitional Probability (FTP)

Sequences	Native speakers		L2 learners	
	% agreement	Cohen's κ	% agreement	Cohen's κ
FTP ≥ .98 (*n* = 11)	85.3	.72	65.9	.41
FTP ≥ .71 (*n* = 33)	69.5	.60	58.7	.46
FTP < .70 (*n* = 67)	21.8	.19	19.6	.17
FTP < .14 (*n* = 11)	11.8	.09	6.0	.05

Discussion

Taken together, findings demonstrate the usefulness of transitional probability for corpus-driven identification of FSs. Forward transitional probability was the sole significant predictor accounting for double-digit proportion of variance in native speakers' and L2 learners' responses. Higher transitional probability values were also linked to greater consistency in terms of the range of responses provided, while lower values corresponded to decreased consistency, again for both native speakers and L2 learners. These findings suggest that transitional probability can be used outside word segmentation research to reveal insights into how words pattern together to form units of meaning, thereby improving the identification of FSs in corpora with the view of using such psychologically valid sequences in language classrooms.

Implications for Methodology and Theory

In previous corpus-driven research, there has been a tendency to identify structures purely according to their frequency of occurrence, applying the range criterion to ensure that the sequences are not restricted to idiosyncratic tendencies of individual users (e.g., Biber & Barbieri, 2007; Cortes, 2004). However, the current findings suggested that the use of frequency as a measure of formulaic status, with a priori decisions regarding sequence length, often fails to accurately identify FSs. A directional measure of word association, transitional probability in fact emerged as a more accurate metric of FSs, insofar as formulaic status can be estimated in a sequence completion task. Compared to mutual information, which provides a general measure of association strength across the entire sequence, and frequency, which reveals how often a particular structure recurs, transitional probability estimates strength of association at crucial points in the structure, leading to more accurate identification of utterance boundaries. Because transitional probability is sensitive to position-specific, directional information, it also

contributes to the detection of more functionally complete sequences that no longer cross semantic and syntactic boundaries, thus reducing the incidence of overlapping and partially repeated structures.

The application of transitional probability to formulaic language research holds potential benefits for those attempting to categorize and describe the presence of FSs in various genres and registers. With the traditional lexical bundle approach often misidentifying FSs and producing lists of overlapping structures that lack functional salience, the standard practice of assigning functions to these sequences becomes a challenge that results in the inconsistent assignment of functional roles across studies, such as, for example, treating *the fact that the* and *the nature of the* as discourse organizing bundles or as serving stance and referential purposes, depending on the particular analysis conducted (e.g., Biber et al., 2004; Cortes, 2004). With the lexical bundles identified in many studies crossing syntactic and semantic lines, it is not surprising the consistent assignment of functional roles proves to be a difficult task. By making use of transitional probability in the identification of FSs, we can create lists of structures that are more functionally salient and use these results to better categorize and describe the language present in the corpora under investigation. For example, high FTPs for the *extent to which* and *in the wake of* indicated relatively stable (and likely formulaic) structures. On the other hand, although highly frequent, low FTP and BTP for *but it is not* indicated unlikely formulaic status.

In the dataset used for this study, forward transitional probability, compared to backward probability, was better at predicting language users' performance. This is unsurprising because completions focused on the last element in each sequence, which was targeted by forward probability. Although backward probability may not have been particularly helpful here, it will likely be as effective for predicting formulaic status of sequences to be completed with the first

word of each sequence deleted. For instance, in the sequence *and there is no*, the low backward probability of .06 suggests that *and* is not part of this structure. On the other hand, in *as soon as possible*, high backward probability of .99 implies that the structure is indeed a four-word sequence. One remaining issue pertains to establishing a threshold of transitional probability to determine a structure's formulaic status. As yet there are no standards; however, current analyses suggest a possible benchmark of .70, which was the transitional probability value associated with the sequences that elicited moderate levels of native speaker agreement in sequence completion (see Table 6).

The current results also provide evidence in support of usage-based models of language (e.g., Barlow & Kemmer, 2000), which posit that the structures with higher transitional probabilities likely represent usage events that have come to be used by language users as multiword units. In fact, native speakers showed high consistency in their completions of sequences with high forward transitional probabilities (≥ .98), suggesting that these sequences might represent units of meaning which have achieved formulaic status. Conversely, sequences with relatively random responses were linked to low transitional probability (< .14), implying that these lack functional salience and therefore do not represent usage events in the minds of users. With native speakers and L2 learners demonstrating similar response patterns, frequency of exposure to the target language and sensitivity to frequency-based, statistical regularities in linguistic input (as indexed through transitional probability) emerge as important variables determining language users' ability to complete target sequences. The finding that L2 learners' responses were predictably less accurate and more constrained in their range than the responses of native speakers, likely reflects L2 learners' less extensive and intensive exposure to English, compared to linguistic experience of native speakers (see Ellis, 2012). As active L2 learners,

especially in the academic domain, the L2 users were likely still in the process of accruing probability statistics needed for them to enjoy the processing benefits of FSs in comprehension and production (e.g., Pawley & Syder, 1987; Tremblay et al., 2011).

Implications for Teaching

The use of transitional probability in future corpus-driven research has the potential to produce more pedagogically valuable sequences that possess greater functional salience, resulting in structures that will be better suited for pedagogical purposes since they should be easier for students to understand and eventually use in their own discourse. Because previous research focusing on the teaching of FSs has yielded mixed results (e.g., Boers et al., 2006; Cortes, 2006), it would seem that which kinds of FSs are being taught, and how, becomes a crucial factor. Even with appropriate instructional techniques, if teachers and students target misidentified sequences, they are unlikely to achieve success. With FSs representative of the language EAP learners are likely to face in their academic careers failing to appear in any meaningful way in many of the most popular EAP texts (Wood & Appel, 2014), the lack of emphasis placed on this aspect of language may partially be due to the fact that these sequences are often misidentified in the literature and are therefore perceived as lacking pedagogical value.

The current research highlights a potential value that FSs hold for teaching and suggests that it might be worth using transitional probability to identify functionally usable FSs in a variety of genres and registers, with the idea of ultimately incorporating them in L2 instructional materials. If this is to take place, specific corpora that focus on the registers and genres most relevant to particular groups of learners (e.g., university-level students) will need to be compiled and used as the source texts for the creation of pedagogically relevant FS lists. This is one area where corpus-driven approaches to the identification of FSs may play an important role since it

is often difficult to accurately identify important formulaic sequences present in specific contexts through personal reflection alone. This is especially true in ESP contexts, where the language instructor or materials creator may not be well versed in the specific type of English that needs to be taught. By using the corpus-driven methods described in the present study, researchers can objectively identify formulaic sequences that will hold the most benefit to the language learner in these contexts.

Limitations and Conclusion

Several limitations of this study must be addressed in future research. First, although the target corpus used in this study was based on British English, all participants were users of North American English. Although attempts were made to control for this limitation, this mismatch in English dialects may have had an effect on at least some participants' ability to complete certain sequences with the target word. Second, the sequence completion task focused exclusively on frequently occurring four-word sequences. The effectiveness of transitional probability to predict sequence completion rates for structures of different lengths remains to be investigated. Finally, transitional probability as a metric of FSs was evaluated in a sequence completion task which likely requires language users to access metalinguistic knowledge about language, instead of targeting the kinds of frequency- and usage-based processing implied by input-driven statistics. Therefore, in future studies, transitional probability must be evaluated in tasks which involve a processing speed component (e.g., timed reading or speaking). Despite these limitations, this research points to an advancement in corpus-based identification of FSs, with the use of transitional probability for creating lists of FSs that are more psychologically valid and salient than those identified using traditional methods. To make the most of corpus-based methods,

regardless of the specific metric used, future research needs to apply these methods to a variety of corpora, with the goal of developing practical, pedagogic solutions for helping L2 learners across a range of settings.

Note

1. Range was also operationalized as a raw value, corresponding to the total number of response options provided, and more conservatively as the number of response options attested in at least 10 participants' data. In all cases, analyses yielded identical findings.

Connecting Study 1 to Study 2

Study 1 demonstrated how aligning theory and method in corpus-informed research can lead to improved approaches and more valuable results. While linking theory and method proved beneficial to corpus-informed formulaic sequence extraction, there are other areas of linguistic inquiry that continue to suffer from issues of methodological rigour. Identification of L1 related production tendencies is one example that would benefit from methodological improvement. In Study 2, a corpus approach to the manual identification of L1 related production tendencies is used a way to explore issues of methodological rigour and demonstrate how increased attention to this issue can help improve the value and applicability of findings in this area.

Chapter 3: Study 2

Linking adverbials in L2 English academic writing: L1 differences

To be submitted to the International Journal of Corpus Linguistics

By Randy Appel

Abstract

Appropriate and effective use of linking adverbials (e.g., *furthermore*, *in addition*, *on the other hand*) plays an important role in discourse cohesion. In the present study, a learner corpus of 150 argumentative essays was examined to determine how linking adverbials were used by L2 English academic writers from three different language backgrounds (Arabic, Chinese, French). Applying contrastive interlanguage analysis, several unique production tendencies related to specific linking adverbials, as well as broader functional categories, were revealed in the writing of each L1 group. Findings included overuse of additive linking adverbials (e.g., *in addition*, *also*) by L1 Arabic writers, contrastive linking adverbials (e.g., however, on the other hand) by L1 Chinese writers, and appositional linking adverbials (e.g., *in fact*, *indeed*) by L1 French writers of L2 English. Methodological and pedagogical implications of these findings are discussed.

Within studies of second language (L2) English academic writing, the use of linking adverbials (e.g., *on the other hand*, *furthermore*, *in contrast*) has received considerable attention due to the important role these items play in structuring text and improving cohesion. In corpus-informed studies on this subject, scholars have highlighted the difficulty L2 learners face when attempting to use linking adverbials in their academic writing, as well as specific patterns of overuse, underuse, and misuse that distinguish L2 writers from their first language (L1) counterparts (e.g., Carrio-Pastor, 2013; Chen, 2006; Crewe, 1990; Field & Yip, 1992; Lei, 2012; Milton & Tsang, 1993; Yeung, 2009). However, due to the prevailing focus on native and non-native writer contrasts, we still lack a clear understanding of which production tendencies are common to all L2 English learners and which may be more specifically associated with L2 English users of particular L1 backgrounds. A distinction between these two factors is important since it can lead to more specialized instruction that better targets the unique challenges each group of learners face when attempting to create cohesion in their L2 English writing. Additionally, research of this kind can provide new insights into the preferred discourse organizing conventions used by each L1 group. To begin addressing these issues, the current study examined a closely controlled corpus of L2 English academic essays to identify unique linking adverbial production tendencies associated with writers from three distinct language backgrounds (Arabic, Chinese, French). Applying a contrastive interlanguage analysis (Granger, 1996), the assembled corpus was used to document linking adverbial production tendencies in the L2 English writing of each L1 group.

Linking Adverbials

Although a wide range of names (e.g., discourse markers/connectives, logical connectors, connective adverbs) have been applied in previous research on this subject, the current study

follows Liu (2008) in adopting the term linking adverbial as a broad and inclusive label to cover a range of single and multiword lexical uints that function as a semantic link between discourse of varying lengths (e.g., clause, sentence, paragraph). Examples of linking adverbials include *indeed*, *in addition*, and *on the other hand*. Linking adverbials commonly occur in sentence initial position, yet they can also appear within sentences, at the beginning of clauses, or separated from the rest of a text via bracketed commas (for a detailed description, see Liu, 2008).

As linking adverbials function in a purely semantic role, they are distinguished from conjunctions which serve as both a syntactic and semantic link between clauses. Thus, while conjunctions cannot be removed from the sentences in which they are found without altering grammatical acceptability, removal of linking adverbials has no impact on grammaticality. Broadly defined, linking adverbials represent a lexical category of English that L2 users frequently struggle with as examples of misuse, overuse, and underuse are frequently identified (Crewe, 1990; Silva, 1993; Yeung, 2009). One likely reason for the difficulty L2 users experience in regards to linking adverbials is that appropriate use requires the ability to identify areas of potential ambiguity (i.e., where discourse relations cannot already be inferred from the text), and discriminately apply appropriate linking adverbials to make these relations clear (Altenberg & Tapper, 1998). Thus, overuse, particularly when marking associations that can already be inferred, may result in reduced readability and increased reader frustration (Crewe, 1990). In contrast, underuse of linking adverbials can lead to confusion regarding how to appropriately interpret a given piece of writing.

As linking adverbials are considered important signposts that demonstrate how writers aim to structure their ideas (Leech & Svartvick, 1994), research on this topic can indicate how in-text cohesion is created, as well as the specific kinds of discourse relations writers wish to

emphasize. As a result, research on this subject carries important pedagogical implications that may lead to more targeted instructional materials that can better address difficulties L2 learners face when writing academic English essays. In response to the important role linking adverbials play in structuring text and improving cohesion, as well as the difficulty L2 learners face with regards to appropriate and effective use, a growing body of research has emerged that aims to investigate how linking adverbials are used in L2 English writing.

Research on Linking Adverbials in L2 English Writing

At least as far back as Altenberg and Tapper (1998) there have been calls for increased research on cohesive devices in L1 and L2 writing. Since this time, numerous scholars have attempted to examine how writers from various L1 backgrounds make use of linking adverbials in their L2 writing, often in comparison to L1 writers of the same target language. By far the most common group of L2 users targeted in these studies have been L1 Chinese English as a Foreign Language (EFL) students (e.g., Bolton et al., 2002; Chen, 2006; Field & Yip, 1992; Lei, 2012; Milton & Tsang, 1993; Yeung, 2009). For example, Milton and Tsang (1993) compared EFL writing of L1 Chinese students in Hong Kong to writing from the Brown (Francis, 1964) and London Oslo/Bergen (Johansson, 1978) corpora, as well as a relatively small corpus of computer science textbooks. Results indicated that a wide range of the 25 single-word linking adverbials in their study were overused by L1 Chinese EFL writers, with the highest rates of overuse related to *lastly*, *besides*, *moreover*, *secondly*, *firstly*, and *consequently*. Similarly, Bolton et al. (2002) compared compositions by L1 Chinese EFL students with a corpus of published academic English writing. Frequency ratio comparisons revealed a general overuse of linking adverbials and other cohesive devices by L1 Chinese students, with the five most overused items being *so*, *and*, *also*, *thus*, and *but*. More recently, frequency ratio comparisons were used to

compare academic writing by MA EFL students in Taiwan (Chen, 2006) and PhD EFL students in mainland China (Lei, 2012), with collections of published academic articles in English. In each study, the authors identified the frequent misuse of *besides*, as well as *what's more*, as typical features of L1 Chinese EFL writers. In the only ESL based study targeting L1 Chinese learners of English, Leedham and Cai (2013) demonstrated once again that L1 Chinese undergraduate students overused *besides* and *what's more* relative to their L1 English counterparts.

While the majority of studies conducted to date have focused on L1 Chinese EFL writers, a comparatively small amount of research has targeted alternative L1 groups, such as French (Granger & Tyson, 1996), Swedish (Altenberg & Tapper, 1998), Spanish (Carrio-Pastor, 2013; Martinez, 2002), Arabic (Modhish, 2012), and Persian (Jalilifar, 2008). For example, Granger and Tyson (1996), in their comparison of L1 French EFL writing from the International Corpus of Learner English (ICLE) and L1 English writing from the Louvain Corpus of Native Essay Writing (LOCNESS), found that L1 French EFL writers overused several linking adverbials (e.g., *indeed*, *for instance*, *moreover*). Although these results were primarily based on native/non-native English writer comparisons, post-hoc contrasts with L1 German EFL writers from the ICLE were also used to confirm results. In terms of L1 Arabic writers of L2 English, in a corpus of 50 EFL academic essays, Modhish (2012) found that these writers employed a limited range of cohesive devices, with *and*, *also*, *so*, and *but* identified as especially frequent. However, since no comparison group of writers was included, it is difficult to interpret the importance of these findings.

In sum, previous research on linking adverbials in L2 English academic writing has largely focused on comparisons between L1 Chinese EFL writers and L1 English users, with

limited research targeting alternative L1 groups. Findings from these studies have identified several examples of overuse, underuse, and misuse related to specific groups of L2 English writers. However, as discussed below, it remains to be seen whether identified production tendencies are unique to the specific L1 groups targeted in each study or common to L2 English learners of all linguistic backgrounds.

L1 Specific and Universal Features of L2 Writing

Prior research on linking adverbials in L2 English academic writing has revealed several patterns of overuse, underuse, and misuse related to ESL/EFL writers of selected language backgrounds. However, due to a common focus on native/non-native contrasts, as well as frequent issues regarding corpus comparability, we still lack a firm understanding of the distinction between L1 specific and universal production tendencies. A distinction between these two factors is important since it can lead to more targeted instructional materials that can better help specific L1 groups identify and improve potential areas of difficulty they are likely to experience when writing in their target L2.

The first major issue that has limited our understanding of universal and L1 specific production tendencies is lack of corpus comparability. Although comparability of corpora should be considered a crucial factor in writer contrasts, studies of linking adverbials in L2 English writing have largely failed to control for important differences in target language proficiency, writing conditions, writing type, and sample length – all of which may impact linking adverbial production (e.g., Chen, 2006; Milton & Tsang, 1993). As a result, it is unclear whether production tendencies identified in these studies are a result of differences in target language proficiency, L1, essay genre, or any other uncontrolled factors. For example, Milton and Tsang (1993) primarily relied on the Brown and LOB corpora as a source of comparison for their

collection of L2 English academic writing. Since the Brown and LOB are both composed of general English, and largely lack the inclusion of any significant amount of academic writing, their use represents a genre incongruence that may have led to a misidentification of L1 based production tendencies. Similarly, Chen (2006), who compared EFL master's papers (e.g., diary, literature review, research proposal, instructional paper) with published academic English research articles is an additional example of mismatched corpora. Although in this case the two corpora were more closely matched for genre, differences in writing type, sample length, and writing sophistication between the novice researchers represented in the L2 corpus, and seasoned academic professionals represented in the comparison corpus, likely led to additional production tendencies that distinguished between these two groups.

A second major issue that has limited our understanding of how L2 English learners vary in their use of linking adverbials is the reliance on one-to-one comparisons of native and non-native writer groups. Since studies following this approach generally aim to contrast L1 English users with L2 English writers of a single linguistic background (e.g., Bolton et al., 2002; Carrio-Pastor, 2013; Chen, 2006; Liu, 2013), they are unable to differentiate between production tendencies attributable to specific L1 groups and those common to all L2 English learners. This is because, as only one L1 group is targeted, there is no way of verifying usage tendencies in reference to other L2 users. For this reason, it may be necessary to forgo the use of native English writers as a benchmark for comparisons, and instead adopt a focus on interlanguage comparisons involving L2 writers of multiple language backgrounds (Ortega, 2011). Unfortunately, in the few studies that have analyzed how linking adverbials are used by L2 English writers of multiple L1s (e.g., Altenberg & Tapper, 1998; Granger & Tyson, 1996), albeit in a post-hoc fashion, issues of corpus comparability continue to limit the validity of findings.

This is largely due to a reliance on analyses involving the ICLE. As writing included in the ICLE is taken from post-secondary institutions worldwide, where target language proficiency, writing conditions, and access to reference materials can all vary widely from one institution, or country, to the next, it is difficult to accurately attribute findings to any single factor. In terms of target language proficiency, although ICLE writers are commonly described simply as 'advanced' (e.g., Gilquin, 2008), the analysis of only a small selection of ICLE essays revealed a proficiency range spanning intermediate to advanced levels (Granger, 2004). With this finding suggesting a uniform proficiency descriptor for the ICLE is inappropriate, caution should be taken when using this corpus to identify production tendencies attributable to specific groups of L2 English writers.

Finally, it should also be noted that limited participant numbers in many of these studies have served to reduce the validity of findings, thereby calling into question their conclusions. For example, Bolton et al. (2002), Carrio-Pastor (2013), Chen (2006), and Lei (2012) have all relied on analyses of extremely limited numbers of participants, with a maximum of 20 L2 English writers in each study. With such small sample sizes, it is unclear how representative these findings are of the L2 populations from which they were drawn.

The Current Study

In light of limitations associated with existing research, and the need for a better understanding of how linking adverbials are used by L2 English writers of diverse L1 backgrounds, the current study used a specially designed corpus of ESL argumentative essays to identify within group tendencies and between group differences in linking adverbial production of writers from three different L1 backgrounds (Arabic, Chinese, French). The selection of these L1 groups was largely based on convenience sampling as these writers represented the most common L2 English users studying at the English medium university from which the essays

were collected. Due to the high frequency with which students from these three L1s register in academic programs at this university, it was hoped that results from this study would lead to more applicable findings the could benefit a large number of L2 English learners. Additionally, as each of these L1s were also targeted in previous studies, their selection allowed for greater comparability with prior EFL based research.

The main research questions guiding this study were: Does the use of linking adverbials in the ESL academic writing of L1 Chinese, French, and Arabic students differ? If so, in which ways? Answers to these questions were expected to provide a greater understanding of how L2 English writers of various linguistic backgrounds structure their academic discourse.

Method

Corpora

The main corpora used in this study were composed of essays from multiple sections of an English for Academic Purposes (EAP) writing course at a large English-medium university in North America. The L2 English learners (50 from each L1 group) were all undergraduate students who had partially met the minimum requirements for entry into the university (75 to 89 on the TOEFL iBT or 6.0 to 6.5 on the IELTS), yet were required to take this EAP writing course due to results of an internal placement test. This course, which uses a content based approach to the instruction of academic English writing, assesses students via three integrated-writing exams written in weeks five (summary), ten (cause & effect), and thirteen (argumentative). The current study focused on the final writing exam, a timed, 3-hour argumentative essay that requires students to write an approximately 500-word essay in response to a given prompt. This task represents a pass/fail final exam. Only those students who received a

passing grade were included in the corpora as this was seen as providing a relatively narrow proficiency range that would allow for between-group comparisons.

At least two weeks prior to the final exam students were given a list of reading articles related to the chosen topic. Using standardized templates, students were instructed to take notes on each of these readings for later reference during the exam. At the beginning of the exam period instructors reviewed these notes to confirm adherence to the template and ensure no additional materials were included (e.g., pre-written passages). All essays were hand-written under identical conditions which allowed for the use of paper dictionaries, thesauri, and student notes on assigned readings.

Since student writing comes from multiple sections of the EAP course in question, the corpus contains responses to multiple essay prompts. However, because these prompts are based on the same general pool of available readings, they share a common focus on issues of economic inequality, with specific topics related to arguments for or against the use of various methods of alleviating economic disparity (e.g., charity, microcredit, government action). Thus, although individual prompts vary, the general subject matter of these essays remains consistent across sections. Because this study focuses on the analysis of linking adverbials, which are relatively syntax and content independent, it is unlikely that prompt variation would have a noticeable impact on reported results.

Before carrying out any analyses, all essays were converted to digital word processing files for ease of use. Corpus statistics for the three corpora can be found in Table 7. In terms of word counts, L1 Arabic students tended to produce the shortest argumentative essays, while L1 French students produced, on average, the longest essays of any writer group.

Table 7

Corpus Statistics

	L1 Arabic	L1 Chinese	L1 French
Running words	28,531	29,354	30,194
Mean essay length	571	587	604
Sentences	1,378	1,522	1,494
Mean sentence length	20.70	19.29	20.21

Extraction of Linking Adverbials

As Bolton et al. (2002) have noted, the identification of linking adverbials is "neither uncontroversial nor finite", and basing extraction on predetermined lists can lead to results that fail to adequately capture the full range of linking adverbials present in the corpora being investigated. Therefore, although existing compilations of linking adverbials and other closely related constructs were reviewed prior to beginning analysis (e.g., Celce-Murcia & Larsen-Freeman, 1999; Liu, 2008; Quirk, Greenbaum, Leech, & Svartvik, 1985), automatic extraction based on pre-existing word lists was not used in the present study. Instead, manual extraction, based on careful reading of all essays was considered preferable.

In extracting linking adverbials from each corpus, two main criteria were used as guiding principles. First, linking adverbials were required to be syntactically and semantically independent units (i.e., could be removed from the sentence in which they were found without significantly altering meaning or grammatical acceptability of the utterance). Second, linking adverbials were required to display evidence of discourse cohesion by linking text together at the clause level or higher. In other words, linking adverbials were required to link clauses, sentences,

groups of sentences, paragraphs, or groups of paragraphs together by helping the reader interpret information that follows in light of what had already been presented. Syntactically independent units that did not serve discourse connecting roles (e.g., *in my opinion*, *nowadays*) were not included. Following manual extraction of all linking adverbials by the main researcher, two research assistants reviewed concordance lines for each item to verify findings. A complete list of linking adverbials identified in the three corpora is provided in Appendix B**.**

Analysis

Because linking adverbials primarily serve as connections between sentences, this study followed Bolton et al. (2002) in using the sentence as the base unit for frequency ratio comparisons. Thus, in all tables frequencies are listed on the basis of average occurrences per 1,000 sentences. In order to avoid attributing frequently occurring, yet idiosyncratic production tendencies of a limited number of writers to any L1 group, only those linking adverbials that appeared in the writing of at least 5 different users (10%) from at least one of the three L1 corpora were highlighted for further analysis.

Upon extracting all linking adverbials from the three corpora, each item was assigned a functional role based on the taxonomy introduced by Quirk et al. (1985). However, as the resultative and inferential categories were considered largely overlapping, inferential and resultative categories were combined in this study. Thus, six broad groupings were used for functional analyses: listing (e.g., *furthermore, first*), summative (e.g., *in conclusion, to conclude*), appositional (e.g., *for instance, in fact*), resultative (e.g., *therefore, as a result*), contrastive (e.g., *in contrast, on the other hand*), and transitional (e.g., *by the way, besides*). Additionally, listing devices were further divided into additive (e.g., *in addition, furthermore*)

and enumerative (e.g., *first, finally*) subcategories to better distinguish between these two functional roles.

Definitions of Overuse and Underuse

As both Chen (2006) and Lei (2012) have noted, a generally accepted method of identifying overuse of linking adverbials in L2 English academic writing has yet to emerge. While frequency ratio analyses are the most commonly used approach (e.g., Altenberg & Tapper, 1998; Bolton et al., 2002; Chen, 2006; Granger & Tyson, 1996; Lei, 2012), criteria vary widely from study to study. Due to the arbitrary nature of definitions, the current study took a relatively conservative approach that primarily focused on the identification of significant differences in functional category production tendencies for each L1 group.

This approach relied on a two step process involving frequency ratio comparisons and one-way ANOVAs. First, to reduce the number of one-way ANOVAs that would be conducted, and thereby limit necessary adjustments to the alpha for significance level testing, functional category frequency ratios were reviewed. To highlight unique production tendencies associated with each L1 group, functional categories with a minimum ±15 occurrences per 1,000 sentences frequency ratio discrepancy was selected for further analysis. Because the present study involved writing from three groups of writers, this criterion was applied in relation to frequencies from both remaining corpora. Thus, for a functional category to qualify for significance testing, a minimum ±15 frequency ratio difference when compared to each of the remaining two groups of L2 English writers was required. In all cases where frequency ratio differences met this criterion, one-way ANOVAs with post-hoc comparisons (where appropriate) were used to assess significance.

To better explain significant functional category differences, individual examples of overuse/underuse related to specific linking adverbials were also extracted. Since this process targeted individual items, as opposed to broader functional categories, a smaller minimum frequency ratio difference was applied. Thus, individual linking adverbial overuse/underuse was based on a minimum frequency ratio discrepancy of ±5 occurrences per 1,000 sentences[2]. As a relatively large number of individual items could be identified as overused/underused by each L1 group, one-way ANOVAs were not used to test for significance; instead, frequency ratio discrepancies served as the sole criterion for individual linking adverbial overuse/underuse. This decision was based on the fact that, due to the large number of analyses that would be required, alpha adjustments would result in an overly restrictive significance threshold. As examples of overuse and underuse are based on comparisons with other groups of L2 English writers, these findings are purely descriptive.

Results

General Findings

Based on the aforementioned criteria, a total of 1,444 linking adverbial tokens of 30 types were identified for further analysis (i.e., had appeared in at least 5 different essays from at least one of the three L1 corpora). Of this total, 436, 459, and 549 came from the L1 Arabic, Chinese, and French corpora, respectively. Based on a frequency ratio per 1,000 sentences, this equates to 316.4 occurrences for L1 Arabic, 301.6 for L1 Chinese, and 367.5 for L1 French ESL writers. Thus, L1 Chinese ESL writers were the least frequent users of linking adverbials with an average production density of one occurrence every 3.3 sentences; L1 Arabic ESL writers averaged a similar density with one occurrence every 3.2 sentences; and L1 French ESL writers were the most frequent users of linking adverbials, with an average of one occurrence every 2.7 sentences.

As a first step in exploring production tendencies in the writing of each L1 group, the top 10 most frequently occurring linking adverbials in each corpus were extracted. As can be seen in Table 2, each group of writers was heavily reliant on a limited range of linking adverbials, with 67-71% of total use in each corpus accounted for by the top 10 most frequently occurring items. While each group of writers demonstrated a preference for a restricted number of cohesive devices, this tendency was most pronounced among L1 Chinese ESL writers (71% coverage for the top 10 items). The high frequency of *however* is especially important to this finding, as this item was responsible for 22% of all linking adverbial occurrences in the L1 Chinese corpus—a higher percentage than any other single item in this study. Although intergroup variation can clearly be seen in Table 8, five items (*however*, *therefore*, *for example*, *moreover*, *thus*) were common to all groups of ESL writers.

Table 8

Top 10 most frequently occurring linking adverbials

L1 Arabic		L1 Chinese		L1 French	
Linking Adverbial	Frequency	Linking Adverbial	Frequency	Linking Adverbial	Frequency
however	47.2 (15%)	however	67.0 (22%)	however	50.2 (14%)
therefore	37.7 (12%)	therefore	34.8 (12%)	in fact	32.1 (9%)
for example	23.2 (7%)	thus	19.7 (7%)	therefore	29.5 (8%)
also	20.2 (6%)	for example	18.4 (6%)	for example	23.4 (6%)
moreover	18.1 (6%)	moreover	17.1 (6%)	indeed	22.8 (6%)
in conclusion	16.7 (5%)	in conclusion	12.5 (4%)	thus	21.4 (6%)

in addition	16.7 (5%)	nevertheless	12.5 (4%)	moreover	20.7 (6%)
in fact	13.8 (4%)	first (of all)	11.8 (4%)	first (of all)	18.7 (5%)
on the other hand	12.3 (4%)	as a result	9.9 (3%)	for instance	16.1 (4%)
thus	10.9 (3%)	furthermore	9.2 (3%)	finally	12.7 (3%)

Functional Category Differences

To more effectively explore linking adverbial production tendencies in the writing of each L1 group and better identify between group differences, functional category analyses using the taxonomy introduced by Quirk et al. (1985) were implemented. Table 9 summarizes functional category frequency ratios for all linking adverbials identified in each of the three corpora (frequency ratios for individual linking adverbials in each functional category can be found in Appendix B).

Table 9

Linking adverbials by functional category

Functional Category	L1 Arabic	L1 Chinese	L1 French
Listing	91.4 (28.9%)	68.2 (22.6%)	97.1 (26.4%)
Additive	73.9 (23.4%)	42.1 (14.0%)	47.5 (12.9%)
Enumerative	17.5 (5.5%)	26.1 (8.7%)	49.6 (13.5%)
Summative	21.1 (6.7%)	14.5 (4.8%)	16.1 (4.4%)
Appositional	52.3 (16.5%)	40.7 (13.5%)	99.8 (27.2%)
Resultative	75.4 (23.8%)	82.2 (27.3%)	74.8 (20.4%)

Contrastive	74.7 (23.6%)	92.7 (30.7%)	77.7 (21.1%)
Transitional	1.5 (0.5%)	3.3 (1.1%)	2.0 (0.5%)
Total	316.4 (100%)	301.6 (100%)	367.5 (100%)

Based on a minimum ±15 frequency ratio discrepancy, five examples of functional category overuse/underuse were highlighted for significance testing. From the L1 Arabic corpus, the additive subcategory of listing devices (+26.4) met the frequency ratio criterion. For L1 Chinese ESL writers, frequency ratios indicated underuse of listing (-23.2), and overuse of contrastive (+15) functional categories. Finally, in terms of L1 French ESL writers, the enumerative (+23.5) and appositional (+47.5) functional categories also indicated potential overuse.

To test whether identified functional category differences were statistically significant, one-way ANOVAs were conducted on each of these five functional categories (additive, contrastive, appositional, enumerative, listing). Due to the high number of tests being run, the alpha for significance testing was set at .01. In each case, histograms and descriptive statistics indicated relative symmetry across all L1 groups and functional categories, as well as an acceptable amount of between group variance (i.e., the largest variance in each case was less than four times that of the smallest). As a result, one-way ANOVAs were considered a valid procedure (Howell, 2013). Results indicated significant differences for four of the five functional categories (Table 10).

Table 10

One-Way ANOVAs for Functional Category Differences

Functional Category	*df*	*F*	*Sig.*	ω
Additive	2	7.07	.001*	.29
Contrastive	2	9.37	.000*	.32
Appositional	2	16.53	.000*	.41
Enumerative	2	8.23	.000*	.30
Listing	2	1.40	.251	-

Note. *$p < .01$

Since ANOVAs for the additive, contrastive, appositional, and enumerative functional categories revealed significant between group differences, post-hoc tests using a Bonferroni adjustment were conducted. For additive linking adverbials, significant differences were found between L1 Arabic and L1 Chinese ($p < .004$), and L1 Arabic and L1 French ($p < .006$). Thus, overuse of additive linking adverbials by L1 Arabic ESL writers was confirmed. For the contrastive functional category, significant differences were identified between L1 Chinese and L1 Arabic ($p < .01$), and L1 Chinese and L1 French ($p < .001$). Therefore, overuse of contrastive linking adverbials by L1 Chinese ESL writers was also confirmed. In terms of appositional linking adverbials, significant differences were found between L1 French and L1 Arabic ($p < .001$), and L1 French and L1 Chinese ($p < .001$). As a result, overuse of appositional linking adverbials by L1 French ESL writers was found to be a statistically significant finding. Lastly, for enumerative linking adverbials, post-hoc contrasts indicated only one significant difference between L1 French and L1 Arabic ($p < .001$). Thus, while enumerative linking adverbial production tendencies did indicate a statistically significant deviation between L1 French and L1 Arabic

ESL writers in this study, this finding was not considered to represent a unique production tendency that could be associated with a single L1 group. In sum, one-way ANOVAs and post hoc comparisons confirmed three of the five functional category differences highlighted through frequency ratio comparisons represented unique L1 related production tendencies (additive, contrastive, appositional).

Overuse and Underuse of Specific Linking Adverbials

To better explain functional category differences and highlight production tendencies for individual linking adverbials associated with writers of particular L1 backgrounds, individual items within each functional category items were reviewed. In total, 11 examples of overuse and 5 examples of underuse were identified. In terms of L1 Arabic writers, two specific examples of overuse were found within the additive subcategory of listing devices: *also* (+12.9) and *in addition* (+8.8). For L1 Arabic ESL writers, underuse of the enumerative functional category was partially explained by relatively rare use of *first/first of all* (–6.9). For L1 Chinese ESL writers, a tendency toward the heavy use of *however* (+16.8) helped explain their overuse of contrastive devices , and one example of underuse, *in fact* (–5.3) was found. Within the L1 French corpus three specific examples of overuse from the appositional category could be identified: *for instance* (+8.1), *in fact* (+18.3), and *indeed* (+17.5). Additionally, three examples of overuse from the enumerative category of listing devices were also found: *finally* (+7.6), *first (of all)* (+6.9), and *secondly* (+5.5).

Discussion

The current study explored the use of linking adverbials among L2 English writers of three distinct L1 backgrounds (Arabic, Chinese, French). In contrast to previous research which has largely targeted native/non-native writer comparisons (e.g., Carrio-Pastor, 2013; Chen,

2006), this study used a contrastive interlanguage approach, with carefully matched corpora, to better identify unique production tendencies in the L2 writing of each L1 group. Due to substantial between group variance in average essay length, frequency ratios (based on average occurrences per 1,000 sentences) were used to more effectively identify unique production tendencies among each L1 group. To assess the significance functional category frequency ratio differences, one-way ANOVAs with post hoc comparisons were run. Results indicated unique functional category overuse related to each L1 group. Finally, frequency ratio discrepancies for individual linking adverbials were reviewed as a way of determining which specific linking adverbials might account for significant functional category differences.

L1 Arabic

In terms of L1 Arabic ESL writers, additive linking adverbials were identified as overused, with two specific examples (*also*, *in addition*) helping to explain this finding. The relative overuse of the additive subcategory, as well as these two items in particular, supports results from Modhish (2012) who found a similar group of cohesive devices (e.g., *and*, *also*, *in addition*) frequently occurred in the 50 L1 Arabic EFL essays he analyzed. While Modhish failed to include a comparison corpus to verify results, the contrastive interlanguage analysis implemented here indicates that overuse of these items can indeed be seen as a feature associated with this particular group of L2 English writers.

This relative overuse of additive linking adverbials highlights an important contrast in the way each group of ESL writers in this study made use of linking adverbials to help structure their argumentative essays, and therefore the kinds of relationships they chose to emphasize. For example, while additive linking adverbials comprised 81% of all listing devices used by L1 Arabic ESL writers, this subcategory represented only 61% and 49% of listing devices used by

L1 Chinese and L1 French ESL writers in this study. Thus, relative to other L1 groups, L1 Arabic ESL writers tended to avoid ordinal relationships in favour of a more additive approach to explicit markers of discourse cohesion. Overuse of this functional category may indicate a preference for an argumentative style that relies on addition as a way of supporting the writer's position.

To better understand production tendencies for additive linking adverbials in L1 Arabic ESL writing, key word searches were used to examine specific instances of overused items from this category. In general, these reviews indicated a potentially problematic tendency, as additive linking adverbials frequently appeared with extremely high levels of concentration in individual paragraphs, often in adjacent sentences. The three examples listed below help illustrate this point.

- **Also**, they offer many services and education to poor families helping them enhance their lives. **Also**, micro finance institutions provide loans at very low interests to business, being an important part in eliminating poverty among many societies around the world.
- **Also** Yunus' bank has lent over $6 billion to 6.6 million people since 1976. Last, several thousand organizations are doing microlending in the developing world. **Also**, it has even been imported into poor areas of the US.
- **And** sadly the world has faced death situations which has been caused by different means of poverty, like lack of food or lack of clothes. Looking at the bright side of the story, many people are trying to pull the poor out of his scary life. **In addition**, many organizations were established to support this cause. **Furthermore**, many individuals already introduced such a great idea to help the poor.

In each of these examples a heavy reliance on the additive functional subcategory is clearly evident as these items represent the majority of linking adverbial occurrences in each passage. This focus on one particular functional category, which represented 24% of total linking adverbial occurrences in the L1 Arabic corpus, may result in reader frustration when repeated over the course of an entire piece of writing (Crewe, 1990). To help remedy this issue, it may be necessary to provide L1 Arabic ESL writers with specialized instruction to help raise awareness of their overreliance on additive linking adverbials, potential issues resulting from this overuse, and how to better incorporate a wider range of cohesive devices in their argumentative essays. In terms of potential reasons for the general overuse of additive linking adverbials among L1 Arabic ESL writers, previous research has suggested that unsuccessful translation attempts may be at least partially to blame. For example, previous research has indicated that L1 Arabic students often mistranslate Arabic linking adverbials *fa* and *oumma* as either *and* or *also* (Saeed & Fareh, 2006; Tahaineh & Tafish, 2011). As *fa* and *oumma* are both considered highly frequent in Arabic writing (Fareh, 1998; Tahaineh & Tafish, 2011), the repeated use of additive linking adverbials may result from unsuccessful attempts to adopt L1 usage patterns. However, more research using alternative corpora is needed to confirm these results.

L1 Chinese

For L1 Chinese ESL writers in this study, contrastive devices were identified as overused. In fact, with 31% of total linking adverbial occurrences, contrastive devices were the most heavily used functional category in the L1 Chinese corpus. The high frequency of *however* (+16.8) helps to account for this finding, as this item represented 72% of all instances in this category. The overuse of contrastive linking adverbials indicates that L1 Chinese ESL writers

tended to emphasize explicit markers of contrast as a way of increasing cohesion. Furthermore, the high frequency of *however* suggests that these writers favoured a point-counter point argumentative style in their academic English writing.

Considering the wealth of existing research targeting L1 Chinese EFL writers, it is somewhat surprising that this is the first time overuse of *however* has been identified. One potential reason for the uniqueness of this finding relates to the fact that previous research on this topic has targeted native/non-native English writer comparisons (e.g., Bolton et al., 2002; Leedham & Cai, 2013). Because *however* is already highly frequent in academic English writing (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Liu, 2008), these comparisons may have prevented overuse of this linking adverbial among L1 Chinese writers from being identified. Thus, overuse of *however* identified in the current study may actually place L1 Chinese ESL writers more in line with observed production tendencies in L1 English academic writing (at least when compared to L1 Arabic and L1 French ESL writers). To better evaluate how this item was used in the L1 Chinese corpus, specific instances were reviewed. Despite occasional misuse, production tendencies did not reveal any specific pattern that would suggest an area of concern.

While *however* was the only individual example of overuse that could be identified in the L1 Chinese corpus, the absence of production tendencies highlighted in previous research involving L1 Chinese EFL writers is also noteworthy. For instance, *what's more*, which has frequently appeared as an overused and misused item (e.g., Chen, 2006; Lee & Chen, 2009; Lei, 2012) did not appear with sufficient frequency in the L1 Chinese corpus, or any other L1 corpus, to qualify for analysis. Similarly, *besides*, which has also been highlighted as a frequently occurring and often misused linking adverbial among L1 Chinese EFL writers (e.g., Chen, 2006; Lee & Chen, 2009; Lei, 2012, Milton & Tsang, 1993; Yeung, 2009) appeared relatively

infrequently in the present study and did not qualify for overuse relative to the other two L1 groups. As a result, difference in setting between the present research (ESL) and past research (EFL) should be considered a potentially influential factor. As home country language teaching pedagogy and instructional materials have been identified as a possible reason for the repeated use of linking adverbials among L1 Chinese EFL writers (Leedham & Cai, 2013), the movement to an ESL environment, which made use of alternative language teaching materials and pedagogy, may have helped mitigate previously acquired production tendencies.[3]

Although no longitudinal data is available, the lack of occurrences for *what's more* and *besides* in the present study, which are generally considered more colloquial than academic (Carter & McCarthy, 2006; Parrott, 2000), may point to a general movement away from previously acquired inappropriate production tendencies in EFL settings toward a closer adherence to academic English writing conventions. As the essays for this study were all written at the end of an intensive (6 class-hours per week) academic English writing course, this period may have provided sufficient exposure and guidance regarding appropriate academic English writing conventions to begin reconfiguring students' inventory of existing linking adverbial production tendencies to be more genre and register appropriate. In this sense, results from the present study may support findings from Leedham and Cai (2013) who found L1 Chinese ESL writers decreased usage for *what's more* and *besides* between years 1 and 2, and year 3.

L1 French

In terms of functional category differences, L1 French ESL writers were found to overuse appositional linking adverbials. The overuse of appositional devices (+47.5), which was also identified by Granger & Tyson (1996) in their analysis on L1 French EFL writing, represents the largest frequency ratio discrepancy, and largest effect size, of any functional category in this

study. Three specific examples of overuse, *for instance* (+8.1), *in fact* (+18.3), and *indeed* (+17.5), helped to account for this finding. In general, overuse of the appositional functional category indicates that L1 French ESL writers prioritize expansion, restatement, and the use of examples as ways of presenting arguments and improving cohesion in their L2 English argumentative essays.

As with overused linking adverbials in other L1 corpora, key word searches were used to explore specific instances. Based on these reviews, appositional linking adverbials were found to be frequently repeated, often in adjacent sentences, or with heavily condensed usage patterns in individual paragraphs. Therefore, as with overuse of additive linking adverbials in the L1 Arabic corpus, this production tendency indicates an overreliance on this functional category which could lead to reader frustration when applied throughout the course of an entire essay. The following three examples are representative of this tendency:

- **In fact**, organizations such as microcredit banks, which do not borrow money to poor because they do not have enough money to ensure it, allow people in need to rise by lending them microloans. **Indeed**, in opposite to formal banks, these institution provide money to poor people. Also, as these institutions target essentially women, they allow women empowerment. **For instance**, each year, the Grameen bank interest return to its clients among who 97% are women.
- **In fact**, in today's society, many are poor. 7 out of 10 people live in countries where income inequalities have increased last 30 years and only 1% of the world global population have almost half (46%) of the world wealth. **Indeed**, according to the World Economic Forum, economic disparities will be the second greatest worldwide

issue in the future. **In fact**, this creates debates about solutions to alleviate poverty in the world.

- **In fact**, hunger is a result of extreme poverty that affects the ability of individuals to escape poverty, causes long-term damage to health and reduce capacity for physical activity. The World Bank group tends to eliminate hunger by creating jobs and new opportunities, **for instance**. Yet, the focus has to be on health care and education. **Indeed**, children represents our future

As demonstrated in these three examples, L1 French writers of L2 English may require further instruction to help raise awareness concerning their overreliance on items from this functional category. Moreover, instruction targeting how to incorporate a wider range of linking adverbials that can be used to emphasize alternative cohesive relationships and better support their argumentative position may also be necessary.

Implications

In previous studies of linking adverbials in L2 English academic writing, researchers have often failed to control for language proficiency differences, relied solely on native/non-native writer comparisons, or analyzed mismatched corpora. The current findings indicate that it is important to supplement these efforts with additional studies using more stringent controls so that we can more accurately identify unique production tendencies in the discourse of each L1 group. Moreover, this study highlights the importance of analyzing L2 discourse produced by writers from multiple L1s, since this has resulted in the discovery of several unique production tendencies not found in previous research (e.g., overuse of contrastive linking adverbials in L1 Chinese ESL writing).

In terms of pedagogical implications, findings from this study reinforce previous claims regarding the difficulty L2 learners face when it comes to appropriate and effective use of linking adverbials in academic English writing (e.g., Crewe, 1990; Yeung, 2009). As indicated by each L1 group's heavy reliance on a limited number of linking adverbials, as well as several examples of functional category and individual item overuse, it would seem that L2 English writers often lack sufficient knowledge of how to effectively use linking adverbials to improve cohesion in their academic English writing. Although various methods of teaching these structures should be introduced and assessed in future research, early results from studies targeting data driven learning (DDL) approaches to this issue are promising (e.g., Boulton, 2009; Cotos, 2014). For example, Boulton (2009) demonstrated that concordance lines for linking adverbials taken from an L1 corpus were a more effective way of teaching linking adverbials to L2 English undergraduates than traditional methods (dictionaries and grammar manuals). Extending this line of research, Cotos (2014) has shown that the addition of L2 corpora for between group contrasts in DDL can lead to increased student engagement and improved learner outcomes. With many publicly available L1 and L2 corpora now easily accessible and searchable, teachers may wish to introduce these tools to their students so that they can better explore appropriate L1 usage tendencies, identify potential problem areas, and further align their L2 writing with the target genre and register they are attempting to acquire.

In addition to the use of DDL to help students better understand and acquire English linking adverbials, lists of frequently mistranslated and misused items may also benefit L2 learners. Furthermore, in-class contrasts of essays that make effective use of linking adverbials, as well as those whose linking adverbial placement does not contribute to cohesion, could help L2 learners better understand the value of linking adverbials in academic English writing

(Yeung, 2009). Ideally, such comparisons should center on authentic genre and register appropriate materials that students will likely encounter in future studies. To avoid frequent repetition of a limited range of linking adverbials, explicit instruction of alternative methods of connecting discourse should also be taught (e.g., references chains, coordinating conjunctions).

While materials and pedagogy targeting particular L1 groups may be necessary to help address specific areas of concern highlighted in this study, it is also important to keep in mind the many similarities across L1 groups. For example, five of the top ten most frequently used items in each corpus were common to all L2 English writers in this study (*however*, *therefore*, *for example*, *moreover*, *thus*), and only four functional categories indicated significant between group differences (additive, contrastive, appositional, enumerative, listing). Thus, while targeted instruction may be necessary to address specific L1 related areas of concern, more universal approaches to the instruction of linking adverbials is also likely to benefit L2 learners from a wide range of L1 backgrounds.

Limitations and Future Research

There are several limitations that should be addressed in future research. First, corpus size and total number of participants limit the applicability of findings. While the number of participants in the present study was sufficient to reveal general trends, and compares favourably with many similarly focused studies, more robust numbers are needed to provide more definite conclusions regarding each of the unique production tendencies identified. Second, future research may wish to move beyond broad L1 descriptors (e.g., Arabic, Chinese, French) toward more specific labels that more accurately identify the specific language background and variety of each participant group (e.g., Cantonese, Mandarin, Parisian French). Third, to generate a more complete picture of how linking adverbials are used in various kinds of academic English

writing, additional writing types (e.g., cause & effect, expository) should also be targeted. Finally, the contrastive interlanguage analysis implemented in the current study enabled the identification of several unique linking adverbial production tendencies among each group of L1 English writers, yet future research is needed to investigate potential underlying causes of the identified production tendencies (e.g., cross-linguistic influence, language teaching pedagogy, language teaching materials). To achieve this goal, it will be necessary to develop comparable corpora of L1 writing produced by each targeted group, as well as collections of language teaching materials, that can be used to help distinguish between these factors.

Conclusion

Two main conclusions can be drawn from this study. First, even when controlling for variance in target language proficiency, writing conditions, and essay type, there are clear differences in how writers from various L1 backgrounds make use of linking adverbials to explicitly mark cohesion in their L2 English argumentative essays. Second, overuse of specific linking adverbials, as well as broader functional categories, by each L1 group suggests that effective use of linking adverbials remains an area of difficulty for many L2 English learners. Based on these findings, it may be necessary for L2 English instructors to highlight frequently overused linking adverbials by each L1 group so that these learners can better identify and improve aspects of cohesion in their academic English writing.

Notes

2. In liner with Bolton et al. (2002) a per 1,000 frequency ratio was selected since this helped reduce the incidence of extremely low frequency scores and added greater comparability with previous research.

3. Proficiency level differences between ESL and EFL writers may also have contributed to these findings. However, with most EFL based studies failing to include any standardized measures of proficiency, it is difficult to explore this possibility.

Connecting Study 1 and Study 2 to Study 3

Taken together, Studies 1 and 2 of this dissertation demonstrated how an appropriate theoretical framework and increased methodological rigour can lead to improvements in corpus linguistics research. While Studies 1 and 2 demonstrated key benefits associated with the application of each of these elements, the focus thus far has resided in well-established areas of linguistic inquiry that are primarily analyzed through of collections of written discourse. Although analyses of writing have always been popular in corpus linguistics due to the stable nature and relative ease associated with collecting and analyzing discourse in this register, newly introduced technologies and software applications are beginning to offer innovative methods of examining L2 English oral discourse in ways that were previously either too time consuming or labour intensive to be used on a large scale.

Study 3 explored the usefulness of two particular software programs, Coh-metrix 3.0 and VocabProfile, which can be used as tools to better understand differences in perceived linguistic ability. By using these software applications to analyze L2 English speech, Study 3 moved beyond the traditional focus on writing to look at the relatively underexplored area of L2 English oral discourse.

Chapter 4: Study 3

Lexical aspects of comprehensibility and nativeness from the perspective of naïve L1 English raters

To be submitted to Bilingualism: Language and Cognition

By Randy Appel & Pavel Trofimovich

Abstract

This study analyzed the contribution of lexical factors to native-speaking raters' assessments of comprehensibility and nativeness in L2 speech. Using transcribed samples to reduce non-lexical sources of bias, 10 naïve L1 English raters evaluated speech samples from 97 L2 English learners across two tasks (picture description and TOEFL integrated). Subsequently, the 194 transcripts were analyzed through statistical software (e.g., Coh-metrix, VocabProfile) for 32 lexical variables spanning measures of lexical diversity, sophistication, and pattern density. In the picture description task, word length and gerund use helped distinguish comprehensibility from nativeness, with lexical diversity contributing to both. In the cognitively more demanding TOEFL task, lexical correlates of comprehensibility and nativeness were similar, with lexical diversity and lexical sophistication measures contributing to both. These findings are discussed in relation to the acquisition, assessment, and teaching of lexical properties in L2 speech.

Introduction

Although diverse in their specific interests, studies of second language (L2) oral (speaking) performance have largely concentrated on global constructs of L2 speaking ability as evaluated by trained raters using prescribed grading rubrics from high-stakes assessments, such as the Test of English as a Foreign Language and the Chinese Ministry of Education's Test for English Majors (Crossley & McNamara, 2013; Crossley, Salsbury, & McNamara, 2015; Crossley, Salsbury, McNamara, & Jarvis, 2011; Iwashita, Brown, McNamara, & O'Hagan, 2008; Lu, 2012). However, the goal of many L2 speakers, including instructed learners, is to prepare for life outside of the classroom (Derwing & Munro, 2015). Therefore, assessments of spoken ability by naïve raters, compared to trained raters or teachers, might provide a more accurate indication of the level of communicative success L2 speakers could achieve while living and working in L2 communities.

Comprehensibility, defined as perceived ease of understanding and operationalized through scalar ratings, has emerged as a practical and reliable means of capturing naïve raters' impressionistic judgments of L2 speech (e.g., Derwing & Munro, 2015). A focus on comprehensibility is consistent with the idea that what is relevant to communication is understandable speech (i.e., speech that is intelligible and easy to understand to interlocutors), not necessarily nativelike or accent-free production (Derwing & Munro, 2015; Levis, 2005). To date, researchers have mostly targeted measures of L2 fluency (e.g., pausing, articulation rate), as well as segmental and prosodic accuracy (e.g., production of vowels and consonants, word stress, intonation contours), in relation to comprehensibility (i.e., Derwing & Munro, 1997; Derwing, Rossiter, Munro, & Thomson, 2004; Hahn, 2004; Kang, Rubin, & Pickering, 2010; Munro & Derwing, 1999; Trofimovich & Isaacs, 2012). Meanwhile, the role of lexis in naïve

raters' assessments of comprehensibility, as potentially distinct from the role of lexis in assessments of nativeness, has remained largely unexplored. Therefore, the goal of the current study was to examine lexical correlates of comprehensibility and nativeness in naïve raters' evaluations of L2 speech across two tasks (picture description, academic summary) by targeting raters' judgments of these constructs in relation to a comprehensive set of 32 finegrained lexical measures of speech content.

Research on L2 Speech

Since the early 1980s, research focusing on various dimensions of L2 speech has targeted specific subconstructs that contribute most to global evaluations of L2 speaking ability (e.g., global proficiency), often using raters' holistic assessments of grammatical accuracy, lexical variation, and pronunciation, among others (e.g., Adams, 1980; McNamara, 1990). For example, Adams analyzed trained raters' global evaluations of L2 speaking ability in relation to holistic judgments of accent, comprehension, fluency, grammar, and vocabulary from oral interviews. For these raters, factors related to grammar and vocabulary were most closely associated with L2 speakers' overall proficiency scores. Similarly, McNamara found that trained rater evaluations of grammar and expression were the strongest predictors of L2 speakers' overall communicative effectiveness. While these studies revealed several subcomponents of L2 speech contributing to global evaluations of L2 speaking ability by trained raters, the use of holistic judgments of grammar or lexis prevented researchers from making more nuanced conclusions about the specific lexical or morphosyntactic factors most relevant to these assessments.

More recently, computer-aided forms of analysis have allowed researchers to analyze L2 discourse in more finegrained ways that are comparable across studies. For example, Lu (2012) used computer-aided extraction of 26 different lexical measures along three categories (lexical

density, sophistication, and variation) to quantify trained English teachers' evaluations of 408 audio recordings of L2 English from the Spoken English Corpus of Chinese Learners (Wen, Wang, & Liang, 2005), showing that automated measures of lexical variation were correlated with rater assessments of L2 oral ability. Similarly, Crossley and McNamara (2013) used computer-aided extraction of measures targeting lexis, topic development, and delivery to model expert rater evaluations of 244 audio recordings of L2 English speech (see also Crossley et al., 2015; Crossley, Salsbury, & McNamara, 2010; Ginther, Dimova, & Yang, 2010; Iwashita et al., 2008), showing that "automated indices related to word type counts and word frequency predicted 61% of the variance of the human scores of overall speaking proficiency" (p. 171).

While this research has resulted in greater reliability and objectivity in the analysis of L2 speech, the emphasis on evaluations by expert/trained raters or experienced L2 teachers has largely remained (e.g., Crossley & McNamara, 2013; Crossley et al., 2011, 2015; Ginther et al., 2010; Iwashita et al., 2008; Lu, 2012). Unfortunately, this focus on trained raters or experienced language teachers has left open the possibility that the features being attended to are simply a result of the training process used to prepare raters, previous exposure to language teaching pedagogy and theory, or adherence to specific features listed in the grading rubric. In light of these limiting factors, it is necessary to look at additional rater populations—including naïve raters (i.e., raters with no specialized training in linguistics or language teaching)—to more fully understand how L2 speech is perceived by potential interlocutors in target language communities (Koizumi, 2012).

Comprehensibility and Nativeness in L2 Speech

Comprehensibility, with its typical focus on naïve rater evaluations, has emerged as a useful construct in assessments of L2 speech. Based on intuitive evaluations characteristic of the

kinds of impressionistic judgments language users make about their daily experiences with language (Oppenheimer, 2008), comprehensibility represents a relatively narrow construct, compared to overall speaking ability which might vary considerably across contexts. Highlighted as an important aspect of L2 speech and a central concern for L2 learners (Abercrombie, 1949; Derwing & Munro, 2015), comprehensibility is often discussed in reference to the competing ideologies of nativeness (accent-free, nativelike L2 speech) and intelligibility (intelligible L2 speech), with comprehensibility included within the broad definition of intelligibility (Levis, 2005). Arguments in favor of a focus on comprehensibility over nativeness stem from a belief that even heavily accented speech can be considered highly comprehensible (Derwing & Munro, 2015), implying that understandable L2 output is ultimately more important to successful communication.

To date, scholars have primarily examined pronunciation and fluency dimensions of L2 speech that are associated with raters' assessments of comprehensibility and nativeness. With respect to comprehensibility, raters appear to attend to various aspects of L2 speech, including segmentals (Munro & Derwing, 2006), prosody (Kang et al., 2010), fluency (Derwing et al., 2004), grammatical accuracy (Derwing, Rossiter, & Ehrensberger-Dow, 2012), and discourse richness (Crowther, Trofimovich, Isaacs, & Saito, 2015a). In contrast, judgments of nativeness appear to be tied exclusively to segmental and prosodic accuracy (e.g., Munro, Derwing, & Burgess, 2010; Saito, Trofimovich, & Isaacs, 2016a). However, the prevailing focus on pronunciation and fluency aspects of L2 speech, in relation to comprehensibility and nativeness, has limited our understanding of the role lexis plays in rater evaluations of these constructs.

Following the research framework developed in L2 vocabulary research (e.g., Crossley et al., 2015; Lu, 2012), Saito et al. (2015) recently provided initial evidence for possible

associations between lexical content of L2 speech and raters' comprehensibility judgements, using samples from 40 French speakers of L2 English who completed a picture description task. Among 12 lexical measures tapping into various dimensions of L2 speech, four specific dimensions—lexical appropriateness, fluency, variation, and sense relations—were identified as relevant to raters' comprehensibility judgments. Conceptualized as a follow-up to this initial investigation, the current study extended these preliminary findings to a larger sample of L2 speakers (97 L2 speakers from varied language backgrounds and L2 proficiency levels) using two different task conditions varying in degree of complexity (an academic listening/speaking task and a picture description task). More importantly, the current investigation represents the first attempt not only to identify specific lexical contributions to raters' assessments of comprehensibility, but also to determine if comprehensibility can be distinguished from nativeness in terms of various measures of L2 lexis.

The Current Study

Given the need to better understand how naïve raters perceive L2 speech and to identify lexical characteristics of L2 comprehensibility (as distinct from nativeness), this study targeted naïve native (L1) English raters' impressionistic judgments of comprehensibility and nativeness in L2 English speech samples recorded by 97 speakers from multiple language backgrounds. Because task type and topic may influence ratings (Crowther et al., 2015b; Kuiken & Vedder, 2014), this study also focused on L2 speech produced across two different tasks to further explore the role of task type in assessments of comprehensibility and nativeness. The study was guided by three questions:

1. Which lexical factors underlie holistic judgements of L2 English comprehensibility and nativeness as judged by naïve L1 English raters?

2. Are the lexical correlates of comprehensibility and nativeness, as evaluated by naïve L1 English raters, distinct? Put differently, can comprehensibility and nativeness be distinguished in terms of their lexical correlates?
3. Do the lexical correlates of L2 English comprehensibility and nativeness vary according to task type?

Method

Speakers

The L2 participants were 97 speakers (20 female, 77 male) with a mean age of 24.2 years (*SD* = 3.14) from an unpublished corpus of L2 English speech (Isaacs & Trofimovich, 2011). The speakers were all international students in undergraduate (19) and graduate (78) programs at a large English-medium university in Canada. The language backgrounds represented in the corpus included Farsi (20), Hindi (11), Telugu (9), Chinese (10), Bengali (9), Punjabi (6), French (6), Spanish (5), Tamil (5), Arabic (4), Gujarati (3), Marathi (2), Urdu (2), Akan, Kannada, Russian, Malayalam, and Portuguese (1 each). L2 speakers had arrived in Canada at a mean age of 23.5 years (*SD* = 3.90), and participated in the study during their first term of university studies. Speakers reported having learned English for an average of 13.5 years (*SD* = 5.13), and estimated using English 10–100% of the time daily (*M* = 63%). All speakers had recently taken either the TOEFL iBT or IELTS. For the speaking component of each test, mean scores were 21.84 (*SD* = 3.19) for the TOEFL iBT and 6.63 (*SD* = .88) for the IELTS. The overall test scores were 90.58 (*SD* = 8.52) for the TOEFL iBT and 6.86 (*SD* = .68) for the IELTS. Self-reports indicated that speakers represented a range (3–9) of L2 speaking ability (*M* = 6.3), based on a 9-point scale (1 = *extremely poor*, 9 = *extremely proficient*).

Speaking Tasks

All speakers performed two speaking tasks varying in cognitive demand. The first task (picture description) involved a series of eight images depicting an encounter between a male and female traveler who realize they have accidentally exchanged bags after bumping into each other on a street corner (Derwing et al., 2004). After reviewing the images, speakers were given 30 seconds of planning time and 60 seconds to create a narrative describing the series of images. The use of this task allowed for comparability of findings with previous studies targeting this task (e.g., Saito et al., 2015, 2016b). The second task (TOEFL integrated listening/speaking task), adapted from TOEFL preparation materials (Educational Testing Service, 2004), required speakers to listen to a brief lecture and read a short paragraph on the same topic before integrating this information into a coherent response that demonstrated understanding of the subject. After listening to the lecture and reading the paragraph, speakers were allotted 30 seconds of preparation time, followed by 60 seconds to speak. Two task versions were used, featuring two topics (actor/observer effects, social influences on perception), with approximately half of the speakers assigned to each. Further analyses showed no consistent differences between these task versions, so all data across TOEFL integrated task versions were pooled together.

Using Robinson's (2001, 2005) task complexity framework, the picture description and TOEFL integrated tasks were evaluated for differences in complexity based on their features. As the picture description task contained less input (eliciting language constrained by the depicted objects, actions, and relationships), contained fewer elements, and did not call for any receptive skills (listening and reading) or reasoning, compared to the TOEFL integrated task, this task was considered the less cognitively demanding of the two tasks. In contrast, the TOEFL integrated task contained multiple elements (listening and reading components) and required reasoning to

integrate these sources into a coherent response. Following Révész, Michel, and Gilabert (2015), speakers' self-ratings were also used to further substantiate claims of task differences in cognitive demand. Upon completion of audio recordings during initial data collection, all speakers were asked to estimate task difficulty using a 9-point scale (1 = *very easy*, 9 = *very difficult*), with the TOEFL integrated task ($M = 4.39$, $SD = 2.02$) rated as significantly more difficult than the picture description task ($M = 3.28$, $SD = 1.87$), $t(92) = 5.45$, $p < .001$, $r = .49$.

Materials

Following previous work (e.g., Saito et al., 2015), speech samples were transcribed by a trained research assistant and verified for accuracy by another trained coder. The goal of transcription was to remove all nonlexical sources of bias that could influence rater evaluations, such as indications of fluency, prosody, and pronunciation (i.e., *when* pronounced as *ven*, *that* pronounced as *zat*). Therefore, spelling was standardized across samples to avoid any sign of accent or pronunciation, and punctuation was removed since these markers could be interpreted as indications pausing and prosody. Any typographic markers of pausing, such as filled pauses (umms, ahhs) and silences (…), were also removed. False starts, word repetitions, and other disfluency markers were retained. Words that could not be transcribed due to audio quality issues or lack of understanding were indicated by /—/. Since punctuation is a feature of written English, and punctuation markers would have been based on the transcriber's subjective judgment, punctuation (including capitalization) was considered inappropriate and removed. All transcripts were checked to ensure a minimum of 95 words per speaker. Although a cutoff of 100 words for certain lexical analyses (e.g., variation) is preferred (Koizumi & In'nami, 2012), a slightly lower minimum word count was implemented to more accurately capture the full range of linguistic abilities represented by the L2 speakers, several of whom produced samples shorter than 100

words in picture description (n = 7) and TOEFL integrated (n = 6) tasks. The resulting corpora were comparable in total words (15,768 vs. 14,201) and mean sample length (in words) across the two tasks (M = 162, *range* = 95–321 vs. M = 146, *range* = 95–215).

Raters

Raters included 10 untrained, naïve L1 English speakers (6 female, 4 male), all students enrolled in non-linguistics and non-education undergraduate programs at the same English-medium university. All raters (M_{age} = 21.3 years, *range* = 19–24) learned English from childhood, with at least one native English-speaking parent, and reported no previous language teaching experience and no prior courses in applied linguistics or related fields. As students at a large university located in a multicultural urban setting with 16% of the student body comprising international students, the raters were highly familiar with L2 English speech by speakers from various language backgrounds.

Rating Procedure

The 194 transcribed L2 English speech samples (97 from each task) were evaluated for comprehensibility and nativeness by raters during two individual rating sessions of about 1.5 hours each. During each session, raters provided evaluations for one task type (picture description or TOEFL integrated), with half of the raters assessing the picture description task first and the remaining half rating the TOEFL integrated task first. Upon completing a short language background questionnaire, raters were given a brief explanation of the two rated constructs. Comprehensibility was defined as ease of understanding, and nativeness was defined as how closely the language resembled that of a native speaker (training materials and sample on-screen interface are provided in Appendix D). After briefly explaining the target constructs, raters were trained on the MATLAB interface (Yao, Saito, Trofimovich, & Isaacs, 2013) used to

administer the task and record evaluations. To eliminate the effect of topic/task familiarity on rater judgements, all raters were familiarized with the materials used to elicit responses in each task before beginning the corresponding rating session.

Transcribed speech samples were presented individually on screen without time limits in a unique random order for each rater. Below each transcript, there were two free-moving 1,000-point scales for assessing comprehensibility and nativeness, with the negative endpoint (corresponding to the rating of 0) labeled by a frowning face and the positive endpoint (corresponding to the rating of 1,000) labeled by a smiling face. Raters were informed that transcribed speech samples were taken from L2 English speakers with a variety of language backgrounds and linguistic abilities, and therefore encouraged to use the full range of each scale. Before beginning each session, raters were given three practice transcripts to test their understanding of the rating procedure and provide an opportunity to ask questions. These practice samples were taken from existing L2 speech samples that did not meet the required minimum word counts for inclusion, but were modified to ensure a minimum of 95 words per sample. To promote careful reading of each transcript before assigning a score, raters were only able to record their ratings after the text had remained on screen for at least 5 seconds.

Lexical Analysis

Following Lu (2012), and based on Read's (2000) model of lexical richness, vocabulary competence was conceptualized as a multifaceted construct related to three broad dimensions: lexical density, sophistication, and variation[4] (described in detail below). Lexical density was evaluated using various word form incidence scores spanning the categories of connectives, situational model, and pattern density. Lexical sophistication was measured using word formation measures (e.g., word frequency, age of acquisition, familiarity, concreteness,

imageability, meaningfulness) as well as coverage for various frequency bands from preestablished word lists. Finally, lexical variation was evaluated via adjusted measure of lexical diversity. To derive specific measures across these categories, the 194 transcribed L2 English speech samples were analyzed for 17 lexical variables using Coh-metrix 3.0 (Grasesser, McNamara, Louwerse, & Cai, 2004), a computational analysis tool providing various density, sophistication, and variation metrics. An additional 13 lexical sophistication measures came from VocabProfile (Cobb, 2016), which allows for the analysis of texts based on word frequencies in several large scale corpora. Because the TOEFL integrated task made use of both a short lecture and accompanying reading to elicit response, degree of overlap between the reading passage and each L2 sample as well as between the listening lecture and each L2 sample were calculated for both versions of the TOEFL integrated task using TextLex Compare (Cobb, 2016). In total, 30 lexical measures were computed for the picture description task and 32 lexical measures were derived for the TOEFL integrated task.

Coh-metrix Measures

Coh-metrix, which adheres to theoretical frameworks that view comprehension as involving various levels of understanding (Graesser, McNamara, & Kulikowich, 2011), uses characteristics of individual words, sentences, and discourse level connections to evaluate text at multiple levels of analysis (McNamara, Graesser, McCarthy, & Cai, 2014). The 17 lexical measures calculated through Coh-metrix spanned six categories.

- *Mean number of syllables* (descriptive statistics category) was used as a measure of the average word length within each transcribed sample. As word length is one potential indicator of readability, this measure likely reflected some aspects of understanding and

processing ease. Because the data involved transcribed speech, mean number of syllables, rather than character length, was considered a more appropriate estimate of word length.

- *Measure of Textual Lexical Diversity* (MTLD, lexical variation category) was used as an adjusted measure of lexical diversity.[4] In general, higher lexical variation is indicative of less lexical overlap (more unique words). Because greater lexical variation is interpreted by raters as a sign of increased linguistic ability (e.g., Crossley et al., 2014; Lu, 2012), MTLD was considered to contribute in similar ways to raters' L2 speech judgments.
- *Causal connectives* (e.g., because, therefore), *logical connectives* (e.g., and, or), and *additive connectives* (e.g., furthermore, moreover) were recorded for each transcribed speech sample (connectives category). Connectives are an important aspect of cohesion that help bind discourse together, making it easier to process and understand. All connectives scores were incidence scores (averaged to occurrences per 1,000 words).
- *Causal verb frequency* (e.g., hit, move), combined incidence score for *causal verbs and causal particles* (e.g., hit, move, because, in order to), and *verb overlap* (situation model category) were computed as measures of causality underlying the situation model in each speech sample. Based on research from cognitive science and discourse processing, situation model refers to mental representations present within a text that go beyond the surface level of word-by-word comprehension (McNamara et al., 2014). In this sense, situation model refers to the rater's mental representations of the meaning conveyed by each sample. As measures from this category are seen as indicators of level of understanding, they were computed for their potential associations with comprehensibility and nativeness. As with the connectives category, incidence scores for causal verbs and combined incidence score for causal verbs and particles were based on

averaged occurrences per 1,000 words. For the verb overlap measure, Coh-metrix uses WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990) to classify verb categories, calculating it on a point-based system where, if two verbs are found to be synonyms, they are awarded a score of 1; otherwise, a score of 0 is given.

- *Average number of modifiers per noun phrase* and two measures of verb form incidence, relative *frequency of gerunds* (calculated as frequency of verbs ending in *–ing*) and relative frequency of *infinitives* (verbs in unmarked form, such as be, have), served as measures of lexical concentration for various parts of speech (pattern density category). Because increased levels of syntactic complexity are associated with greater processing difficulty (Perfetti, Landi, & Oakhill, 2005), these measures were used as potential indicators of comprehensibility and nativeness.
- *Word frequency*, *word age of acquisition*, *word familiarity*, *word concreteness*, *word imageability*, and *word meaningfulness* (word formation category) were calculated for each sample. Word frequency counts indicate the frequency with which each word appearing in a sample occurs within the English language in general, using the 17.9 million word CELEX corpus (Baayen, Piepenbrock, & Gulikers, 1995) and may help differentiate discourse produced by users of varying linguistic abilities (Crossley et al., 2014; Laufer & Nation, 1995). To arrive at a score for each sample, Coh-metrix uses the logarithm of word frequency for all identified words. Any words that are absent from the CELEX corpus receive a score of 0 and are therefore not included in the resulting score. The remaining measures in this category were computed based on information from the MRC Psycholinguistics Database (Wilson, 1988) which contains scalar ratings from adult English speakers for a wide variety of English words. Word age of acquisition is an

averaged indication of the age at which words are acquired by native speakers. With more complex words being acquired later than simple words, raters may interpret this measure as a sign of linguistic maturity. Word familiarity is an estimate of the average level of familiarity for all words present within each sample to adult English users and may be related to changes in L2 English proficiency (Salsbury, Crossley, & McNamara, 2011). Word concreteness is a representation of the average level of concreteness, or non-abstractness for all words in a sample. As L2 learners tend to acquire concrete words at earlier stages of linguistic development (Crossley, Salsbury, & McNamara, 2009; Ellis & Beaton, 1993), raters may interpret increased word concreteness as an indication of lower linguistic ability. Word imageability is an averaged score representing the ease/difficulty of constructing a mental image for all words in each sample. As with word concreteness, L2 learners tend to learn more imageable words earlier and more easily than less imageable ones (Ellis & Beaton, 1993; Salsbury et al., 2011). Thus, this measure may serve as an additional indicator of L2 ability as assessed by naïve raters. Lastly, word meaningfulness provides an estimate of the level of association among content words in each sample. Less meaningful words (e.g., sine, squib) involve fewer associations, while more meaningful words (e.g., quick, quiet) evoke associations with a wider range of other words. As learners' linguistic ability improves, they begin to use less meaningful words with fewer associations (Salsbury et al., 2011), suggesting that this might impact raters' judgments of comprehensibility and nativeness. For each of these measures, Coh-metrix averages psycholinguistic ratings for all content words present within each sample.

VocabProfile Measures

VocabProfile and TexLex Compare, available through Compleat Lexical Tutor (Cobb, 2016), are online tools used to compare lexis in user-supplied discourse samples with various preestablished word lists and to calculate lexical overlap between two user supplied samples.

- Four separate word lists from the VocabProfile database were used to calculate additional lexical sophistication measures: Browne, Culligan, and Phillips' (2013) New General Service List (NGSL), Browne, Culligan, and Phillips' (2013) New Academic Word List (NAWL), Neufeld and Billuroglu's (2005) Billuroglu-Neufeld List (BNL), and Nation's (2012) British National Corpus and the Corpus of Contemporary American English word lists (BNC/COCA). These four word lists were targeted on the assumption that they might provide unique coverage statistics due to the disparate nature in which each word list was created. With the exception of the NAWL, which provides a single lexical sophistication measure based on the entire 965-word list, all remaining lexical sophistication measures are divided into separate estimates indicating coverage for specific frequency bands. For example, BNL sophistication was divided into measures representing coverage of the three most frequent bands (*BNL_1*, *BNL_2*, *BNL_3*), as well as percentage of words appearing in each transcribed speech sample that did not appear in the BNL (*BNL_Off*). All overlap measures were calculated as percentage of token overlap. As more frequently occurring words are likely to be processed with greater facility, sophistication measures for the most frequently occurring word bands were viewed as a potential indicator of speaking ability and associated comprehensibility and nativeness.

- Overlap between each transcribed sample and the corresponding transcribed lecture (*lecture overlap*) and overlap between each transcribed sample and the reading paragraph (*reading overlap*) were computed, using percentage of token overlap, for each sample from the TOEFL integrated task (lexical overlap category). The working assumption for the inclusion of these variables was that greater overlap with task materials would be perceived as more native and comprehensible.

Results

Comprehensibility and Nativeness Ratings

Interrater reliability for the two constructs evaluated by the 10 raters (Cronbach's alpha) met or exceeded the preestablished benchmark of .70–.80 (Larson-Hall, 2010) for both comprehensibility (α = .79) and nativeness (α = .77) in the picture description task, and for both comprehensibility (α = .81) and nativeness (α = .81) in the TOEFL integrated task. As a result, the 10 individual ratings for comprehensibility and nativeness in each task were averaged to attain a single mean score for each construct in each transcribed speech sample. As expected, correlations between comprehensibility and nativeness ratings were high in both the picture description (r = .93) and TOEFL integrated (r =.94) tasks. However, raters provided overall lower ratings for nativeness than for comprehensibility, in the picture description task (445 vs. 541), $t(96) = 21.02$, $p < .001$, $r = .91$, and in the TOEFL integrated task (440 vs. 536), $t(96) = 20.78$, $p < .001$, $r = .90$. No significant differences in ratings were found across the two tasks.

Lexical Variables and Rated Constructs

As a first step in exploring the role of lexical variables in naïve raters' assessments of comprehensibility and nativeness, Pearson correlation coefficients were computed using all previously described lexical measures (30 for the picture description task, 32 for the TOEFL

integrated task) and the two target constructs, separately per task. Tables 11 and 12 summarize all significant relationships for the two tasks ($\alpha = .05$); a full list of all correlation coefficients in each task can be found in Appendix E.

Table 11

Significant Correlations for the Picture Description Task

Lexical variable	Comprehensibility	Nativeness
Word length	.31*	.30**
MTLD	.31**	.32**
Verb overlap	–.27**	–.24*
Gerunds	–.26*	–.19
Infinitives	.22*	.27**
Word frequency	–.28**	–.25*
Word age of acquisition	–.22*	–.20
Word concreteness	–.24*	–.21*
Word meaningfulness	.23*	.19
NAWL	.24*	.28**
BNL_0	–.23*	–.18

Note. *$p < .05$, **$p < .01$, two-tailed.

In the picture description task (Table 11), 11 lexical variables spanning five of the six targeted categories from Coh-metrix and two lexical sophistication measures from VocabProfile were found to be significantly correlated with comprehensibility. Six lexical variables from the same five targeted categories and one lexical sophistication measure from VocabProfile were significantly correlated with nativeness in this task. While all seven variables correlated with nativeness were also significantly correlated with comprehensibility, the four additional variables

that were uniquely correlated with comprehensibility (gerunds, word age of acquisition, word meaningfulness, BNL_0) helped to distinguish these two constructs.

Table 12

Significant Correlations for the TOEFL Integrated Task

Lexical variable	Comprehensibility	Nativeness
Word length	.29**	.33**
MTLD	.49**	.44**
Additive connectives	–.25*	–.21*
Causal verbs	.17	.22*
Causal verbs & particles	.23*	.18
Modifiers per noun phrase	–.25*	–.22*
Word frequency	–.22*	–.20
Word age of acquisition	.27**	.27**
Word familiarity	–.27**	–.30**
NGSL_3	.31**	.34**
NGSL_Off	–.23*	–.24*
BNL_1	–.17	–.21*
BNL_2	.23*	.27**
BNL_Off	–.33**	–.35**
BNC/COCA_Off	–.34**	–.36**
BNC/COCA _2	.21*	.23*
BNC/COCA _3	.26**	.28**
Lecture overlap	–.21*	–.22*

Note. *$p < .05$, **$p < .01$, two-tailed.

In the TOEFL integrated task (Table 12), there was a wider range of lexical variables associated with comprehensibility and nativeness. This was largely due to the increased number

of lexical sophistication measures from VocabProfile that could be linked to naïve rater assessments of each construct. In fact, while only two lexical sophistication measures from VocabProfile showed significant correlations with either construct in the picture description task, eight such measures were significantly associated in the TOEFL integrated task. For comprehensibility, a total of 16 measures including eight lexical variables covering all six targeted categories from Coh-metrix as well as seven lexical sophistication variables and one lexical overlap measure from VoabProfiler were found to be significantly correlated with this construct. For nativeness, 16 (largely identical) variables showed significant correlations. The lexical variables distinguishing between the two constructs in this task included causal verbs, causal verbs and particles, word frequency, and BNL_1.

Lexical Predictors of Comprehensibility and Nativeness

To better understand the relationship between the two rated constructs and the lexical variables associated with them (see Tables 11 and 12), stepwise multiple regressions were carried out, with comprehensibility and nativeness used as criterion factors. However, a more restrictive inclusion criterion for predictor variables was implemented, given that multiple lexical variables had significant associations with each rated construct. First, a minimum correlation coefficient value of .25 ($p < .05$) was set, as this is considered the benchmark for a small association in L2 research (Plonsky & Oswald, 2014). Second, the relationship between the identified predictor variables was checked for multicollinearity. As in previous research (e.g., Crossley et al., 2011), a typical collinearity threshold of .70 was set. Using this threshold, two predictor variables in the TOEFL integrated task were found to be highly correlated: NGSL_Off and BNC/COCA_Off ($r = .80$, $p < .01$). Therefore, only BNC/COCA_Off was included in the subsequent multiple regressions. As maximum correlation between all other predictors was .53,

collinearity was not considered a problem for the remaining variables. To ensure comparability between target constructs, all predictors meeting the inclusion criteria in a specific task were included in the regressions for both constructs. For the picture description task, seven variables met the inclusion criteria and were therefore included in the multiple regressions for this task (word length, MTLD, verb overlap, gerunds, infinitives, word frequency, NAWL). For the TOEFL integrated task, 10 variables met the criteria (word length, MTLD, additive connectives, modifiers per noun phrase, word age of acquisition, word familiarity, NGSL_3, BNL_2, BNC/COCA_Off, BNC/COCA_3). The results of multiple regression analyses are summarized in Tables 13 and 14.

Table 13

Results of the Multiple Regression for the Picture Description Task

Criterion variable	Predictors	R^2	ΔR^2	*B*	95% CI	*t*	*p*
Comprehensibility	Word length	0.10	0.10	0.446	[0.051, 0.841]	2.24	.027
	MTLD	0.15	0.05	0.003	[0.001, 0.005]	2.48	.015
	Gerunds	0.20	0.05	–0.005	[–0.010, –0.001]	–2.43	.017
Nativeness	MTLD	0.10	0.10	0.003	[0.001, 0.005]	3.09	.003
	NAWL	0.18	0.08	0.033	[0.009, 0.057]	2.70	.008
	Infinitives	0.21	0.03	0.001	[0.000, 0.003]	2.02	.047

In the picture description task (Table 13), comprehensibility was predicted by a combination of word length, MTLD, and gerunds variables, with 20% of total variance explained. Beta values showed that, while both word length and MTLD were positively

associated with comprehensibility, the use of gerunds showed a negative association. Nativeness was predicted by a combination of MTLD, NAWL, and infinitives (21% of total variance explained). Although regression models for both comprehensibility and nativeness contained MTLD as a significant predictor, the remaining two variables were unique to each construct. In contrast, for the TOEFL integrated task (Table 14), both comprehensibility and nativeness were predicted by the same four variables (MTLD, NGSL_3, BNC/COCA_Off, BNL_2), with similar individual contributions and a comparable total variance explained (39% for comprehensibility, 41% for nativeness).

Table 14

Results of the Multiple Regression for the TOEFL Integrated Task

Criterion variable	Predictors	R^2	ΔR^2	*B*	95% CI	*t*	*p*
Comprehensibility	MTLD	0.24	0.24	0.005	[0.003, 0.007]	4.40	.000
	NGSL_3	0.32	0.08	0.026	[0.012, 0.040]	3.70	.000
	BNC/COCA_Off	0.36	0.04	–0.013	[–0.024, –0.002]	–2.36	.020
	BNL_2	0.39	0.04	0.008	[0.001, 0.014]	2.29	.024
Nativeness	MTLD	0.19	0.19	0.004	[0.002, 0.006]	3.58	.001
	NGSL_3	0.30	0.11	0.029	[0.015, 0.042]	4.30	.000
	BNC/COCA_Off	0.36	0.06	–0.016	[–0.026, –0.005]	–2.96	.004
	BNL_2	0.41	0.06	0.009	[0.003, 0.015]	2.89	.005

Lexical Variables as Unique Predictors of Comprehensibility and Nativeness

Because multiple regressions were based on analyses of all transcribed speech samples, regardless of speakers' sample length, language background, or L2 speaking ability (as assessed independently through TOEFL iBT speaking scores), it was important to determine lexical

predictors of comprehensibility and nativeness that are independent from such sample- and speaker-specific factors. Indeed, there were significant associations between sample length, speakers' language background, and speakers' TOEFL-based speaking proficiency and rater judgments of comprehensibility and nativeness (r = .20–.46) and between these sample- and speaker-specific factors and the lexical variables selected for multiple regressions (r = .21–.30). More finegrained analyses of lexical correlates of comprehensibility and nativeness were also warranted because several target lexical measures were likely dependent on sample- and speaker-specific factors, such that, for instance, longer samples, samples produced by speakers of higher speaking proficiency, or those by speakers from certain language backgrounds could contain more instances of connectives or infinitival forms. Therefore, for the final analysis, partial correlations were used to determine relationships between comprehensibility, nativeness, and each of the significant lexical factors from multiple regressions, while controlling for three sample- and speaker-specific factors (*TOEFL speaking score*, *L1 background*, *word count*). Results of partial correlations are summarized in Tables 15 and 16.

Table 15

Partial Correlations for the Picture Description Task

Variable	Comprehensibility	Nativeness
Word length	.25*	—
MTLD	.21*	.22*
Gerund	–.25*	—
NAWL	—	.18
Infinitives	—	.12

Note. Variables partialled out from each correlation included TOEFL iBT speaking score, L1 background (coded categorically), and word count. *p < .05, two-tailed.

In the picture description task (Table 15), all three variables previously identified as significant predictors of comprehensibility maintained significant, yet modest, associations (Plonsky & Oswald, 2014). Thus, word length, MTLD, and gerunds were each important to assessed comprehensibility regardless of L2 speakers' TOEFL speaking ability, L1 background, or sample length. Conversely, for nativeness, partial correlations revealed only one lexical measure (MTLD) that maintained a statistically significant correlation after controlling for text- and speaker-specific factors.

Table 16

Partial Correlations for the TOEFL Integrated Task

Variable	Comprehensibility	Nativeness
MTLD	.53**	.45**
NGSL_3	.19	.22*
BNL_2	.31**	.35**
BNC_COCA_Off	–.36**	–.38**

Note. Variables partialled out from each correlation included TOEFL iBT speaking score, L1 background (coded categorically), and word count. *$p < .05$, **$p < .01$, two-tailed.

In the TOEFL integrated task (Table 16), of the four predictors of comprehensibility, only one (NGSL_3) failed to reach significance after partialling out text- and speaker-specific factors. For nativeness, all four predictor variables maintained significant correlations with the target construct. In most cases, associations were medium in strength (Plonsky & Oswald, 2014), accounting for up to 28% of shared variance.

Discussion

The current study explored the relationship between lexical measures of L2 speech and assessments of L2 comprehensibility and nativeness across two tasks. In contrast to previous research which has targeted trained raters' evaluations of L2 speech (e.g., Crossley et al., 2010; Lu, 2012), this study examined assessments by naïve raters, using transcribed samples as opposed to audio recordings to target specific contributions of lexis to judgments of comprehensibility and nativeness.

Lexical Correlates of Speech Ratings

In response to the first research question, which asked which lexical factors underlie holistic judgements of L2 English comprehensibility and nativeness, a measure of lexical variation (MTLD) was identified as the sole common significant predictor of scores for both constructs in each task. Therefore, lexical variation can be seen as an important element in naïve raters' assessments of L2 comprehensibility and nativeness (at least in English), regardless of task type, which is consistent with prior research demonstrating that lexical variation is a significant factor in L2 English linguistic ability (e.g., Crossley et al., 2014; Saito et al., 2015). As L2 speakers advance through stages of development, their lexicons begin to expand. As a result, more advanced L2 speakers are able to make use of a greater proportion of different word types in their oral productions, and in doing so are able to improve the comprehensibility and nativeness of their speech. Naïve raters (Saito et al., 2015) and trained assessors (Crossley et al., 2014; Lu, 2012) recognize increased lexical variation as a sign of greater linguistic ability, associating it with enhanced comprehensibility and nativeness. The important role of lexical variation is further reinforced through results of the partial correlations which indicated that MTLD was the sole predictor associated with both constructs in each task. Thus, regardless of

sample length, language background, and independently assessed speaking proficiency, lexical variation is a key aspect of L2 English spoken comprehensibility and nativeness, as judged by naïve raters. While MTLD was a significant predictor of naïve rater assessments of comprehensibility and nativeness in each task, all remaining lexical variables identified through multiple regressions and partial correlations were specific to individual tasks, constructs, or both

Comprehensibility Versus Nativeness

For the second research question, which assessed level of independence between comprehensibility and nativeness (with respect to lexical characteristics of L2 output), analyses revealed task differences. In the picture description task, there appeared to be a clear separation in lexical correlates of the two constructs (see Tables 3 and 5). However, in the TOEFL integrated task, raters' assessments of comprehensibility and nativeness were largely indistinguishable, with nearly identical sets of lexical variables accounting for comparable amounts of variance in each construct.

In the picture description task, beyond the common measure of MTLD, significant predictor variables of comprehensibility included word length and gerund use. The significance of word length as an indicator of comprehensibility likely relates to higher ranked samples containing longer, more complex vocabulary that is representative of later stages of linguistic development. Post hoc analyses support this conclusion, with negative correlations identified between average word length and frequency bands for the most commonly occurring words in the NGSL ($r = –.28$, $p < .01$), BNL ($r = –.37$, $p < .01$), and BNC/COCA ($r = –.37$, $p < .01$). Thus, for L2 speakers to be judged as more comprehensible, they must move beyond a reliance on the most frequently occurring vocabulary and begin using rarer items that are generally characterized by longer words acquired in later stages of L2 development. Because word frequency ($r = –.38$, p

< .01) and word concreteness ($r = –.28$, $p < .01$) were also negatively associated with average word length and with comprehensibility (as shown in Table 1), an increase in average word length may be indicative of improved comprehensibility linked to the use of less frequent and more abstract lexis (Crossley et al., 2009; McNamara et al., 2014). The use of more sophisticated vocabulary allows speakers to be more precise, thereby increasing comprehensibility.

The negative correlation between gerund use and comprehensibility highlighted in both multiple regression and correlation analyses also points to a nontrivial association. Although labeled as "gerunds" in Coh-metrix, this measure may be more accurately referred to as an index of *–ing* forms since the Coh-metrix text parser is unable to accurately distinguish between gerunds and participles sharing the same form (McNamara et al., 2014). With the picture description task requiring speakers to create a narrative describing a series of events, repeated use of *–ing* forms may have decreased perceived comprehensibility by obscuring ordering and suggesting several overlapping continuous actions. For example, in the sample text below, which received a high gerund score, sequencing of events is primarily achieved through fronting information regarding the picture being referred to, while the actual language use suggests ongoing events.

- …the first picture is showing a corner of two roads… two people are coming from two different directions those people are intersecting the third picture is showing the collision so they are holding their head he is putting his spectacles on…

In this sample, each image is described in correct order; however, the repeated use of *–ing* verbs suggests a series of overlapping actions. Thus, repeated use of *–ing* forms may have contributed to decreased comprehensibility by obscuring event sequencing. In contrast, the sample text listed

below, which elicited a low gerund score, makes use of more varied tense and aspect forms to describe the same series of images.

- two people are walking the street… they contact each other suddenly… they didn't see each other from the other side after that when they want to continue their way… when they arrive home… when they open their bags…

Here, the speaker employs very few *–ing* forms, essentially limiting their use to the initial description of the first image in the series, relying on other tense and aspect forms for all remaining events in the narrative. This increased use of infinitives, which was also a significant predictor of nativeness in this task, indicates that preference for specific (uninflected) verb forms can influence perceptions of comprehensibility and nativeness by naïve raters.

Similar to comprehensibility, nativeness in the picture description task was linked to three lexical variables (MTLD, NAWL, infinitives). However, after partial correlations were used to control for differences in TOEFL speaking score, language background, and sample length, only MTLD maintained a significant relationship with nativeness, suggesting that the two constructs can be clearly differentiated in this task. As with previous research looking at linguistic correlates of comprehensibility and nativeness in audio recordings of this same picture description task (e.g., Trofimovich & Isaacs, 2012; Saito et al., 2016b), nativeness was associated with a smaller range of linguistic variables than comprehensibility. Thus, nativeness can be considered a more narrowly defined construct than comprehensibility, at least in this task.

In the TOEFL integrated task, lexical differences between comprehensibility and nativeness were less clearcut. In fact, the same four variables (MTLD, NGSL_3, BNL_2, BNC/COCA_Off) emerged as significant predictors in the multiple regressions for both constructs, with comparable total amounts of variance explained. These variables point to the

same common underlying factor identified in the picture description task—notably, rich and varied lexical use. With NGSL_3 and BNL_2 representing indices of lexical use beyond the most frequent bands in each word list, these measures once again indicate the importance of lexical sophistication to assessments of L2 ability. Although the relationship between NGSL_3 and comprehensibility failed to reach significance in the partial correlations, lexical sophistication measures seem to hold promise as future tools in the assessment of L2 discourse. To appropriately implement these measures, it is important to look not only at coverage for the most frequent bands from each list, but also at off-list measures, as BNC/COCA_Off was found to hold a significant negative correlation with scores of comprehensibility and nativeness for the TOEFL integrated task. The value of this measure as an indicator of L2 speaking performance is likely related to its ability to identify nonsense words, lexical inventions, grammatical errors, and false starts contributing to lower comprehensibility and nativeness ratings, such as *whe* as a false start of when, *wha* as a false start for what, and *feeled* instead of felt. As each of these elements would fail to appear within in-list frequency bands, they are relegated to off-list measures, thereby helping identify discourse that could be considered by raters as being less comprehensible and less native.

Task Effects

In response to the third research question, which targeted possible task effects in the identified lexical correlates of comprehensibility and nativeness, results pointed to two observations. The first observation concerned the total variance accounted for by each of the regression models and the importance of lexical sophistication measures in these models. In the picture description task, 20–21% of the total variance was explained for the two target constructs. In the TOEFL integrated task, the total variance explained by each model was

substantially higher, with 39 and 41% of comprehensibility and nativeness scores accounted for by the predictors. The most likely cause of this discrepancy is differences in task complexity. According to the Cognition Hypothesis (Robinson, 2005), cognitively more demanding tasks, compared to simpler tasks, result in more elaborate language with richer and more complex vocabulary and grammar (e.g., Robinson, 2001). Thus, cognitively demanding tasks likely call for the use of more varied and sophisticated vocabulary, increasing the likelihood for communication difficulties to arise, at least with respect to the use of vocabulary, which may have led to greater total variance being explained by lexical measures in the TOEFL integrated task. Conversely, for the picture description task, which arguably elicited simpler language constrained by the visual input, there may not have been as many lexical measures relevant to raters' assessments of comprehensibility and nativeness, which would account for the lower amount of total variance explained and fewer significant associations, particularly with lexical sophistication measures.

The second observation relates to differences in degree of construct independence discovered in the two tasks. While multiple regressions and partial correlations for the picture description task revealed a clear separation between lexical correlates of comprehensibly and nativeness, results from the TOEFL integrated task indicated greater overlap, and therefore increased difficulty in distinguishing between these two constructs. One possible reason for this finding is register-based differences. As with level of complexity, there is a clear difference in terms of the register required to respond to each task. The picture description task, which used a series of illustrated images (depicting a nonacademic scenario) to stimulate a narrative describing an event sequence, is unlikely to be found in any content-based postsecondary program. Alternatively, the TOEFL integrated task, which aims to simulate a common event in a content-

based academic program (i.e., reading and listening to academic discourse before displaying a coherent understanding of the material) requires a higher level of academic English.

Implications

In terms of implications for theory, findings of this study question the scope of distinction (at least in lexical terms) between comprehensibility and nativeness as partially overlapping yet independent constructs of L2 speech. That the two target constructs were largely indistinguishable in a cognitively more complex task eliciting academic language suggests that this distinction is likely task-specific, in the sense that the linguistic dimensions relevant to each construct vary with the linguistic and cognitive demands of a given speaking task. In addition, because the majority of prior evidence for the independence of these constructs came from analyses of audio recordings of L2 speech (e.g., Trofimovich & Isaacs, 2012), conclusions drawn from these studies may have been overtly influenced by pronunciation and fluency related factors (e.g., speech rate, segmental errors). Thus, the distinction between comprehensibility and nativeness appears to be sensitive to the mode in which L2 speech is evaluated, because the use of transcribed samples in this study eliminated potential speech- and fluency-related factors in favor of a purely lexical focus. What emerges, then, is a complex relationship between linguistic correlates of comprehensibility versus nativeness, one that must be situated within task and register differences and identified in relation to both spoken and written features of discourse.

In terms of practical implications, findings suggest that regardless of the desired goal (comprehensibility or nativeness), L2 speakers would benefit from a focus on increasing depth and breadth of vocabulary knowledge, as lexical variation (MTLD) was revealed as an important variable in naïve rater assessments of both constructs in each task. The unique contribution of the current dataset is in demonstrating that the importance of lexical variation in evaluations of L2

speech is relevant to learners from multiple language backgrounds, as opposed to those from particular languages, such as French (Saito et al., 2015; Trofimovich & Isaacs, 2012), Chinese (Lu, 2012), and Japanese (Saito et al., 2016a). Thus, a focus on lexical instruction should benefit all learners. Perhaps the most important implication is that, at least when it comes to more academically oriented tasks, it may not be necessary for teachers and learners to adopt an exclusive focus on either comprehensibility or nativeness, as these constructs were largely overlapping in the TOEFL integrated task. Therefore, teachers and learners can focus on both goals simultaneously when working within the academic register.

Limitations and Future Research

As this study was largely exploratory, there are several limitations that should be addressed in future research. In relation to task differences in assessments of comprehensibility and nativeness, findings were based on the analysis of only two tasks. With the TOEFL integrated task being both more cognitively demanding and featuring a different register than the picture description task, it remains to be seen which of these factors (complexity or register), and to which extent, resulted in the observed task effects. Additionally, despite efforts to analyze speech performances from users with a range of L2 speaking proficiencies, this study may not have captured the full range of desired L2 ability levels. Because the L2 speakers were mainly graduate students with moderate to advanced levels of L2 speaking proficiency (as assessed through the speaking component of the TOEFL iBT and IELTS), future research should target a wider range educational levels and language proficiencies, ideally controlling for potentially influencing factors, such as age of acquisition (Moyer, 2013), length of residence (Derwing & Munro, 2013), and aptitude (Granena, 2014).

Furthermore, because the naïve raters were all undergraduate students, they may have lacked sufficient familiarity with academic English to distinguish between perceptions of comprehensibility and nativeness in this register. With their exposure to academic English relatively limited, the concept of academic English—in terms of its comprehensibility, nativeness, or both—may not yet have fully developed. Thus, future research on assessments of comprehensibility and nativeness in L2 academic discourse may wish to target more academically experienced raters (i.e., postgraduate students). Furthermore, additional rater populations, such as L2 English users should also be targeted as this may also lead to new insights on rater perceptions of L2 English speech. Lastly, the use of Coh-metrix may also be interpreted as a potential limitation due to the algorithms used to derive scores for several measures. For instance, in reference to gerund use, Coh-metrix may lack sufficient precision. As the Coh-metrix text parser can result in questionable word class assignments, caution should be taken when interpreting results based on these measures.

Conclusion

Three main conclusions can be drawn from this study. First, as a substantial portion of variance in naïve rater assessments of comprehensibility and nativeness was attributed to lexical (and associated grammatical) features, variables targeting these aspects should be regarded as an important indicator of L2 spoken ability. Second, with each task being associated with a separate set of significantly associated predictors (beyond the common measure of MTLD), it can be concluded that task type and register hold a strong influence on linguistic correlates of comprehensibility and nativeness. Finally, construct independence for comprehensibility and nativeness appears to depend on task type (complex/academic vs. simple/nonacademic), the mode in which spoken discourse is evaluated (spoken vs. written), and also likely on raters'

experience with the target language domain (academic vs. nonacademic). Given these findings, researchers might wish to refine, and perhaps even leave aside, the debate about comprehensibility and nativeness in academic tasks, at least from a lexical point of view, as these constructs (albeit in this dataset) were seen as indistinguishable in naïve raters' assessments.

Notes

4. Although Read's original model includes a fourth dimension (number of errors in lexical use), we follow Lu (2012) in focusing exclusively on the first three dimensions of lexical richness.

5. Lexical diversity and lexical variation are often used interchangeably. However, because the label lexical diversity can have different meanings in different studies (Nation & Webb, 2011), we have used the more precise term lexical variation to refer to the proportion of different word types used in a text.

Chapter 5: Conclusion

Introduction

The studies in this dissertation represent individual, yet interlinked, attempts to improve our understanding of SLA by looking at the lexical characteristics of L1 and L2 English discourse from a corpus-informed perspective. While corpus-informed approaches to linguistic inquiry have gained popularity over the last several decades, it is important that we do not become complacent in the methodologies we employ, but consistently strive to push the field forward by introducing new techniques and improved approaches. The three studies presented here contribute to this goal by demonstrating how corpus-informed methods can be improved upon by establishing a closer adherence to established theoretical frameworks, using increased methodological rigour, and applying new statistical techniques. In doing so, this dissertation targeted three key areas important to SLA that are well suited to this methodology: extraction of formulaic sequences from large scale corpora, identification of unique L1 related production tendencies, and the quantification of lexical features correlated with assessments of L2 English speech performance. These three topics are important to SLA researchers since they inform our general understanding of L1 and L2 acquisition. While each study has made its own unique contributions, with specific objectives, methods, and results, they were also designed to work together as a cohesive whole. In this chapter, I begin by providing a brief summary of each manuscript which focuses on key findings and the important links that bind these studies together. Next, I discuss insights resulting from this set of studies and what it adds to our collective body of knowledge regarding SLA. Following this, pedagogical implications, as well as limitations and directions for future research are provided. Finally, I close with some brief concluding remarks.

Overview of Key Findings

Study 1 was primarily concerned with the first objective of this dissertation: improving methods of extracting formulaic sequences from large scale corpora. With the frequent lack of a theoretical basis in previous research of this kind leading to lists of incomplete, overlapping, or overly extended structures that lack psycholinguistic validity and pedagogical usefulness (Liu, 2012; Nekrasova, 2009; Simpson-Vlach & Ellis, 2010), a first step in Study 1 was to adopt an appropriate theoretical framework that could be used to ground the research. In this case, it was usage-based, or emergentist, models of language. Basing the identification of formulaic sequences within this framework resulted in a necessary shift in focus from simple recurrent word combinations to form function pairings that represent valid units of meaning (e.g., constructions). To identify these form function pairings, Study 1 evaluated transitional probability as a possible metric of unit status. Results revealed that the application of transitional probability reduced many problems associated with traditional approaches to formulaic sequence extraction, leading to lists of items that were more functionally salient and psycholinguistically valid. Key findings from Study 1 were (a) transitional probability can be used to improve the psycholinguistic status, and therefore pedagogical usefulness of formulaic sequences extracted from large scale corpora, and (b) the application of an appropriate theoretical framework can lead to improvements in corpus informed research of this kind.

The methodological focus of Study 1, which targeted improvements in the identification of fixed-form formulaic sequences, also carries over to Study 2. However, in the case of Study 2, attention was placed on improving methods of identifying L1 related production tendencies in the use of linking adverbials by L2 English academic writers. This goal was selected because existing research on this subject has largely relied on methods that fail to control for potentially

influencing factors, such as differences in target language proficiency, writing conditions, and essay genre. To overcome these limitations, Study 2 used a specially designed corpus of ESL essays that had been carefully controlled for each of these features. Employing a contrastive interlanguage analysis that targeted learners from three different language backgrounds (Arabic, Chinese, French), the collected corpus was analyzed for unique production tendencies in the use of linking adverbial by each L1 group. Key findings from Study 2 included (a) the identification of several unique production tendencies not found in previous research and (b) a better understanding of the relationship between L1 background and L2 production.

Together, Study 1 and Study 2 highlight the importance of continuously improving corpus-informed methods of linguistic inquiry so that findings carry increased validity. To achieve this goal, these studies focused on well-established areas of corpus research that are primarily analyzed via collections of written discourse. Although analyses of writing have always been popular in corpus linguistics due to the stable nature and relative ease of analysis, new technologies and software applications are beginning to offer innovative ways of examining L2 speech that were previously either too time consuming or labour intensive to be used on a large scale. To further extend corpus research, Study 3 moved beyond a focus on writing to look at spoken language from a corpus-informed perspective. More specifically, Study 3 aimed to identify lexical factors indicative of differences in naïve L1 English rater assessments of L2 English spoken ability. Although spoken English is increasingly targeted in SLA research, this is still a growing area with much left to learn, and lexical factors contributing to naïve rater' perceptions of spoken ability remain understudied. To help fill this research gap, Study 3 analyzed a corpus of transcribed speech samples from 97 L2 English speakers, across two tasks, to discover which lexical features contributed most to naïve L1 English raters' evaluations of

comprehensibility and nativeness and identify features that helped distinguish between these two constructs. Key findings from Study 3 included (a) the identification of task based differences in rater assessments of comprehensibility and nativeness and (b) specific lexical measures that helped distinguish between these two constructs.

Conclusions from the Three Studies

Importance of Theory in Corpus-Informed Research

Results from this dissertation demonstrate the value of utilizing theory to explain findings from corpus-informed studies and how a closer relationship between research and theory can lead to improved methods and results. In Study 1, the theoretical underpinning provided by usage-based models resulted in statistically significant improvements to the psycholinguistic status of corpus derived formulaic sequences. The movement to adopt an appropriate theoretical framework was crucial in achieving this result as it helped better situate the research so that more effective methods of identification could be applied. In Study 3 theory was once again used for beneficial purposes, yet in this case its role was in helping inform the selection of which lexical measures to target and better interpreting subsequent findings. For example, in Study 3, the three main dimension of lexical richness (variation, sophistication, and density) proposed by Read (2000) were used as a guide in selecting relevant linguistic measures, and the cognition hypothesis (Robinson, 2005) was presented as a potential explanation for the task based differences that were discovered. The benefits gained through the combination of guidance and interpretation provided by theory leads to the conclusion that, where appropriate, it is important for corpus linguists to better integrate linguistic theories in their research efforts (Gries, 2010).

L1 related differences in L2 Acquisition

The second major conclusion that can be drawn from this body of research is made in reference to Study 2 and the importance of the L1 in SLA. While Study 1 and Study 3 treated L2 English learners as a unified group, regardless of L1 background, Study 2 separated L2 English learners into three groups, in an attempt to identify unique production tendencies associated with L2 English writers of each L1 background. Using interlanguage comparisons of L1 Arabic, Chinese, and French ESL writers, findings indicated that production tendencies for several linking adverbials, as well as broader functional categories, could be linked to differences in L1 background. These findings led to the conclusion that it may be necessary to place greater emphasis on L1 background when attempting to improve L2 learners' target language abilities (at least in relation to the use of linking adverbials). As a result, it is important for teachers and students to recognize the benefits and drawbacks associated with production tendencies of particular L2 users so that these propensities can be addressed in the classroom. Regardless of whether the root cause of these differences is cross-linguistic influence, home country language teaching pedagogy, or home-country language teaching materials, it is clear that the ways writers from different language backgrounds use linking adverbials to structure their L2 English written discourse is not identical. These findings are important to keep in mind when attempting to generalize findings from studies that only target L2 writers from a specific L1.

Task Based Differences in L2 Assessment

Another major conclusion that can be drawn from this dissertation is the impact and importance of task differences in evaluations of L2 English speech. This conclusion, based on results from Study 3, supports findings from variationist SLA literature, which has typically used interviews, ethnographies, and longitudinal data to discover how language varies in relation to

factors such as task type, setting, and social class (e.g., Adamson, 1980; Beebee, 1980; Selinker, 1972). Study 3 extends this research by demonstrating the value of corpus data, and associated automated forms of analysis, in the identification of task differences, particularly in relation to assessments of comprehensibility and nativeness.

While comprehensibility and nativeness are frequently viewed as competing goals in SLA (e.g., Levis, 2005), with resultant theoretical and practical implications, results from Study 3 suggest that it may be necessary to reevaluate this distinction and better contextualize potential differences between these constructs within the speaking task(s) being targeted. For example, while naïve rater evaluations for the picture description task in Study 3 indicated a clear separation between comprehensibility and nativeness, results from the more academically oriented, and cognitively demanding, TOEFL integrated task demonstrated that lexical correlates of comprehensibility and nativeness were largely indistinguishable. As a result, evidence from the TOEFL integrated task brings into question the notion that comprehensibility and nativeness are in fact competing goals, and suggests that a hardline distinction between these constructs may not always be necessary (at least from a purely lexical perspective).

Pedagogical Implications

This research carries several important pedagogical implications for language teachers. First, as results from Study 1 suggest inadequate knowledge of formulaic sequences (FSs) among L2 English undergraduates, it is important that we begin to better incorporate FSs into instructional materials geared toward these learners. While methods of teaching these structures must also be developed and tested, the extraction of appropriate FSs would seem to be a logical first step. Since the extraction of FS by way of transitional probability has been shown to result in more psycholinguistically valid and pedagogically useful structures, it may be worth using the

statistical approach introduced in Study 1 to identify FSs that could then be incorporated in L2 instructional materials. It should however be noted that if researchers and teachers choose to use this approach, these methods must be applied to appropriate corpora made up of texts from genres and registers that are most relevant to the specific group of L2 learners being targeted.

Once genre and register appropriate FSs have been extracted, data driven learning (DDL) and learner data driven (LDD) methods could be used by teachers and students to explore appropriate usage tendencies and identify common problem areas L2 English learners are likely to encounter with regards to specific FSs. For example, after a brief introduction to the user interface, the freely available Contemporary Corpus of American English (COCA) could be used by students to explore discourse functions, frequency of occurrence, distributions patterns, and common contexts of extracted FSs. To take advantage of this tool, teachers could assign short lists of FSs to students that could then be used for keyword searches. After identifying important distinguishing characteristics of each assigned FS, students could incorporate this new-found knowledge into a short presentation that introduces these items to remaining class members. Additionally, this approach could easily be adapted to other relevant linguistic features, such as linking adverbials (Study 2), and task specific lexical correlates of linguistic ability (Study 3).

Alternative DDL activities that could be incorporated into language classrooms include the assignment of search terms based on teacher identified production errors in student writing. For example, teachers noticing consistent problems with appropriate use of specific vocabulary could assign the lexical item in question as a key word to be used in corpus searches. This approach would allow for a more personalized way of helping students identify and remedy deficiencies in their L2 writing. Lastly, where possible, comparisons between L1 and L2 corpora could be also be used to highlight common areas of difficulty that arise among L2 learners and

indicate inappropriate, yet common usage tendencies that require explicit attention (e.g., the overuse of additive linking adverbials among L1 Arabic writers of L2 English). Thus, while this dissertation has primarily presented corpora as tools for researchers, the growing availability of corpora, as well as the ease with which they can be analyzed, suggests that corpus analysis may also prove valuable for L2 learners and teachers.

In addition to recommendations for the use of DDL and LDD methods, results from Study 3 carry important pedagogical implications of a separate nature. First, regardless of desired goal (comprehensibility or nativeness), it would appear that L2 English speakers would benefit from a focus on increasing depth and breadth of vocabulary knowledge, as lexical variation was the sole variable common to rater assessments of both constructs in each task. Importantly, the value of lexical diversity seems to transcend differences in task type, target construct, sample length, and speaking proficiency. Thus, all L2 English speakers should benefit from a focus on improving lexical diversity. Therefore, a direct implication of this research is that vocabulary should play a prominent role in any L2 English classroom. Second, when it comes to more academically oriented tasks, results from Study 3 suggest that it may not be necessary to adopt an exclusive focus on either comprehensibility or nativeness, as these constructs were largely indistinguishable in the TOEFL integrated task. This should come as welcome news to EAP teachers and learners, as what have previously been viewed as competing goals may actually hold the same lexical basis.

Limitations and Suggestions for Future Research

Limitations specific to each study have already been discussed in individual chapters. However, there are also several broader limitations that apply to this dissertation as a whole. First, despite findings from Study 2 that indicate there are important differences in how L2

English users from various L1 backgrounds use L2 English, Study 1 and Study 3 treated L2 English users as a homogenous group. While this was purposeful, and a necessary choice given participant constraints, it may prove valuable for future researchers to extend findings from Study 1 and Study 3 to target specific L1 groups so that potential L1 related differences can be identified. Second, although methodological improvements have been applied as a unifying theme in this dissertation, the collection of studies presented here is not without issues of methodological rigour. No study, including those that compose this dissertation, is without fault and it is therefore necessary for replication studies to retest and reevaluate the results identified here.

In addition to the need for replication studies to help further validate results from this dissertation, new studies are needed to extend each branch of research. For example, as mentioned earlier, the statistical method presented in Study 1 should be used in future research to extract pedagogically valid formulaic sequences from various corpora made of texts L2 learners are likely to encounter in their target language environments. Upon extraction of these sequences, methods of teaching these structures to students should also be researched and tested. This is a subject I am particularly interested in and am eager to begin exploring.

Furthermore, in relation to Study 3, expanding the range of lexical variables to include measures of formulaic language would also present an interesting addition. However, to do so would require the development of new approaches, as current methods were considered insufficient for this purpose. With new software applications and statistical techniques continuously being released, it is likely that more effective methods of measuring the formulaic language used by L2 English learners will be released in the near future. Finally, there is an ongoing need for longitudinal studies targeting L2 English users of various L1s and target

language proficiencies so that we can better understand how these learners evolve and adapt as they spend more time in target language communities. While Study 2 used comparisons with previous EFL based research to hypothesize that time spent in target language environments may have resulted in some of the noticeable differences between EFL and ESL based research, true longitudinal analyses are needed to confirm this hypothesis.

Concluding Remarks

This dissertation is the culmination of several years of work and countless hours of reading, writing, and revising. A long and arduous process, this experience has also been rewarding as each of the topics presented in this dissertation represents a long standing interest. As a result, I am grateful to have had the chance to explore these topics in greater detail. That being said, I am also excited to begin expanding my efforts to address additional topics that drive my interest in SLA.

References

Abercrombie, D. (1949). Teaching Pronunciation. *English Language Teaching, 3*, 113-122.

Abercrombie, D. (1966). *Studies in phonetics and linguistics*. London: Oxford University Press.

Adams, M.L. (1980). Five co-occuring factors in speaking proficiency. In J. Firth (Ed.), *Measuring spoken proficiency* (pp. 1-6). Washington, DC: Georgetown University Press.

Adamson, H. D. (1980). *A study of variable syntactic rules in the interlanguage of Spanish-speaking adults acquiring English as a second language*. Unpublished doctoral dissertation, Georgetown University.

Altenberg, B., & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learners' written English. In. S. Granger (Ed.) *Learner English on computer* (pp. 3-18). New York: Longman.

American Council on the Teaching of Foreign Languages. (1999). ACTFL proficiency guidelines-speaking. Retrieved from http://www.actfl.org/files/public/guidelinesspeak.pdf.

Aslin, R., & Newport, E. (2012) Statistical learning: From acquiring specific items to forming general rules. *Current Directions in Psychological Science, 21,* 170-76.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *CELEX*. Philadelphia, PA: Linguistic Data Consortium.

Bamberg, B. (1983) What makes a text coherent? *College Composition and Com, 34,* 417-29.

Barlow, M. (2011). Corpus linguistics and theoretical linguistics. *International Journal of Corpus Linguistics, 16,* 3-44.

Barlow, B., & Kemmer, S. (2000) Introduction: A usage-based conception of language. In M. Barlow and S. Kemmer (eds.) *Usage Based Models of Language*. Stanford: CSLI

Publications. 1-64.

Beebee, L. (1980). Sociolinguistic variation and style-shifting in second language acquisition. *Language Learning, 30*, 433-447.

Biber, D. (2009) A corpus-driven approach to formulaic language in English.: Multiword-patterns in speech and writing. *International Journal of Corpus Linguistics*, *14,* 275-311. d

Biber, D., & Barbieri, F. (2007) Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26,* 263-86.

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

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

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006) Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research, 10,* 245-61.

Bolton, K., Nelson, G., & Hung, J. (2003). A corpus-based study of connectors in student writing: research from the international corpus of English in Hong Kong (ICE-HK). *International Journal of Corpus Linguistics, 7*, 165-182.

Boulton, A. (2009). Testing the limits of data-driven learning: language proficiency and training. *European Association for Computer Assisted Language Learning, 21*, 37-54.

Browne, C., Culligan, B., & Phillips, J. (2013). The new general service list. Retrieved from http://www.newgeneralservicelist.org.

Carrio-Pastor, M. (2013). A contrastive study of the variation of sentence connectors in academic English. *Journal of English for Academic Purposes*, *12,* 192-202.

Carter, R. & McCarthy, M. (2006). *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

Celce-Murcia, M., & Larsen-Freeman, D. (1999). *The grammar book: An ESL/EFL teacher's course* (2nd ed.). Boston: Heinle & Heinle.

Chen, C. (2006). The use of conjunctive adverbials in the academic papers of advanced Taiwanese EFL learners. *International Journal of Corpus Linguistics, 11*, 113-130.

Cobb, T. (2016). Compleat Lexical Tutor [computer program]. Accessed January 2016 at http://www.lextutor.ca/

Conklin, K., & Schmitt, N. (2008). The processing of formulaic language. *Annual Review of Applied Linguistics, 32,* 45-61.

Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly*, *34*, 548-560.

Cortes, V. (2004) Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23,* 397-423.

Cortes, V. (2006) Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education, 17,* 391-406.

Cotos, E. (2014). Enhancing writing pedagogy with learner corpus data. *European Association for Computer Assisted Language Learning, 26*, 202-224.

Crewe, W.J. (1990). The illogic of logical connectives. *The ELT Journal, 44*, 316-325.

Crossley, S., & McNamara, D. (2013). Applications of text analysis tool for spoken response grading. *Language: Learning & Technology, 17,* 171-192.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2009) Measuring L2 lexical growth using hypernymic relationships. *Language Learning, 59*, 307–334.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning, 60*, 573-605.

Crossley, S. A., Salsbury, T., & Mcnamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics, 36*, 570-590.

Crossley, S., Salsbury, T., McNamara, D., & Jarvis, S. (2011). What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly, 45*, 182-193.

Crowther, D., Trofimovich, P., Saito, K., Isaacs, T. (2015a). Second language comprehensibility revisited: investigating the effects of learner background. *TESOL Quarterly, 49*, 814-837.

Crowther, D., Trofimovich, P., Isaacs,T., & Saito, K. (2015b). Does speaking task affect second language comprehensibility? *Modern Language Journal, 99*, 80-95.

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34*, 5-34.

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching, 42*, 476-490.

Derwing, T., & Munro, M. (1997). Accent, intelligibility, and comprehensibility. *Studies in Second Language Acquisition, 12*, 1–16.

Derwing, T.M., Rossiter, M.J., Munro, M.J. & Thomson, R.I. (2004). L2 fluency: Judgments on different tasks. *Language Learning, 54*, 655-679.

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: evidence-based perspectives for L2 teaching and research*. Philadelphia, PA: John Benjamins Publishing Company.

Derwing, T. M., &Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A seven-year study. *Language Learning, 63*, 163–185.

Educational Testing Services (2004). Independent Speaking Scoring Rubrics. Retrieved from http://www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf

Ellis, N. C. (2012) Frequency-based accounts of second language acquisition. In S.M. Gass and A. Mackey (eds.), *Routledge Handbook of Second Language Acquisition*. New York: Routledge. 193-210.

Ellis, N., & Beaton, A. (1993) Psycholinguistic determinants of foreign language vocabulary acquisition. *Language Learning, 43*, 559–617.

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

Field, Y., & Yip, L.M. (1992). A comparison of internal conjunctive cohesion in the English essay writing of Cantonese speakers and native speakers of English. *RELC Journal, 23*, 15-28.

Firth, J. (1935) The technique of semantics. *Transactions of the Philological Society, 34*, 36-77.

Fletcher, W. (2011) *Phrases in English*. Retrieved from phrasesinenglish.org

Francis, W. N. (1964). *Manual of information to accompany a standard sample of present-day edited American English, for use with digital computers*. Department of Linguistics, Brown University.

Gilquin, G. (2008). Combining contrastive and interlanguage analysis to apprehend transfer: detection, explanation, evaluation. In G. Gilquin, S. Papp, and M.B. Diez-Demar (Eds.). *Linking up contrastive and learner corpus research* (pp. 3-33).

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing, 27,* 379-399.

Graesser, A.C., & McNamara, D.S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science, 3,* 371-398.

Graesser, A.C., McNamara, D.S., Louwerse, M.M., & Cai, Z. (2004). Coh-Metrix: analysis of text on cohesion and language. *Behavioural Research Methods, Instruments, and Computers, 36,* 193-202.

Granena, G. (2014). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning, 63,* 665-703.

Granger, S. (1996). From CA to CIA and back: an integrated contrastive approach to computerized bilingual learner corpora. In K. Aijmer, B. Altenberg & M. Johansson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Learning*. Amsterdam: John Benjamins.

Granger, S. (1998). The computer corpus: a versatile new source of data for SLA research. In. S. Granger (Ed.) *Learner English on computer* (pp. 3-18). New York: Longman.

Granger, S. (2003). The international corpus of learner English: a new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly, 27,* 538-546.

Granger, S., & Thewissen, J. (2005). Towards a reconciliation of "Can Do" and "Can't do" approach to language assessment. Paper presented at the Second Annual Conference of EALTA (European Association of Language Testing and Assessment), Voss, Norway, 2-5 June 2005.

Granger, S., & Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes, 15*, 17-27

Gries, S. (2010). Corpus Linguistics and theoretical linguistics. *International Journal of Corpus Linguistics, 15,* 327-343.

Hahn, L. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*, 201-223.

Halliday, M.A.K. (2005). *Computational and quantitative studies*. London: Continuum.

Halliday, M.A.K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Howell, D.C. (2013). *Statistical methods for psychology*. California: Wadsworth.

Hulstijn, J. H., Schoonen, R., de Jong, N. H., Steinel, M. P., & Florijn, A. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing, 29*, 203-221.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hyland, K. (2008) As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27,* 4-21.

Isaacs, T., & Trofimovich, P. (2011). *International students at Canadian universities: Validating a pedagogically-oriented pronunciation scale.* Unpublished corpus of second language speech.

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility. *Studies in Second Language Acquisition, 34,* 475-505.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: how distinct? *Applied Linguistics, 29,* 29-49.

Jalilifar, A. (2008). Discourse Markers in Composition Writings: The case of Iranian learners of English as a foreign language. *English Language Teaching, 1,* 114-122.

Johansson, S. (1978). *Manual of information to accompany the Lancaster-Oslo/Bergen corpus of British English, for use with digital computers*. Department of English, University Of Oslo.

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of English language learner proficiency in oral English. *Modern Language Journal, 94,* 554-566.

Kaeding, J. (1897). *Haufigkeitworterbuch der deutschen. Sprache, Steglitz*: privately published.

Kennedy, G. (1991). Preferred ways of putting things with implications for language teaching. In J. Svartvik (Ed.) *Directions in corpus linguistics* (pp. 335-373). New York: Mouton de Gruyter.

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: the role of listener experience and semantic context. *The Canadian Modern Language Review, 64,* 459-489.

Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: using short texts with less than 200 tokens. *System, 40,* 554-564.

Kuiken, F., & Vedder, I. (2014). Raters' decision, rating procedures and rating scales. *Language Testing, 31,* 279-284.

Kuiper, K. (1996) *Smooth Talkers*. Mahwah, NJ: Erlbaum.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-74.

Larson–Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written

production. *Applied linguistics, 16*, 307-322.

Lee, D., & Chen, S.X. (2009). Making a bigger deal of the smaller words: function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing, 18*, 149-165.

Leech, G. (1991). Corpora and theories of linguistic performance. In J. Svartvik (Ed.) *Directions in corpus linguistics* (pp. 105-122). New York: Mouton de Gruyter.

Leech, G., & Svartvick, J. (1994). *A communicative Grammar of English*. Longman: London.

Leedham, M., & Cai, G. (2013). Besides…on the other hand: Using a corpus approach to explore the influence of teaching materials on Chinese students' use of linking adverbials. *Journal of Second Language Writing, 22,* 374-389.

Lei, L. (2012). Linking adverbials in academic writing on applied linguistics by Chinese doctoral students. *Journal of English for Academic Purposes, 11,* 267-275.

Levis, J. (2005). Changing contezxts and shifting paradigms in pronounication teaching. *TESOL Quarterly, 39,* 367-377.

Light, R.J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin, 76,* 365-77.

Liu, B. (2013). Effect of first language on the use of English discourse markers by L1 Chinese speakers of English. *Journal of Pragmatics, 45*, 149-172.

Liu, D. (2008). Linking adverbials: an across-register corpus study and its implications. *International Journal of Corpus Linguistics, 13,* 491–518.

Liu, D. (2012). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes, 31,* 25-35.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Review, 96,* 190-208.

Mair, C. (1991). Commentary: the importance of corpus linguistics to understanding the nature of language. In J. Svartvik (Ed.) *Directions in corpus linguistics* (pp. 335-373). New York: Mouton de Gruyter.

Martinez, A. C. L. (2002). The use of discourse markers in E.F.L. learners' writing. *Revista Alicantina de Estudios Ingleses, 15,* 123-132.

McCarthy, P.M., & Jarvis, S. (2010). MTLD, voc-d, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behaviour Research Methods, 42,* 381-392.

McNamara, D., Graesser, A., McCarthy, P., Cai, Z. (2014). *Automated evaluation of text and discourse with coh-metrix*. Cambridge, M.A.: Cambridge University Press.

Meara, P.M., & Bell, H. (2001). P_Lex: a simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect, 16*, 5-24.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K.J. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography, 3,* 235-244.

Milton, J., & Tsang, E.S.C. (1993). A corpus-based study of logical connectors in EFL students' writing: directions for future research. In R. Pemberton & E.S.C. Tsang (Eds.) *Studies in Lexis* (pp. 215-246). Hong Kong: The Hong Kong University of Science and Technology Language Centre.

Mirman, D., Estes, K., & Magnuson, J. (2010). Computational modelling of statistical learning: effects of transitional probability versus frequency and links to word learning. *Infancy, 15,* 471-86.

Modhish, A. (2012). Use of discourse markers in the composition writings of Arab EFL learners. *English Language Teaching, 5*, 56-64.

Moyer, A. (2013). *Foreign accent: The phenomenon of non-native speech*. Cambridge University Press.

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

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: an advanced resource book*. Taylor & Francis.

McEnery, T., & Wilson, A. (2001). *Corpus Linguistics. Edinburgh*: Edinburgh University Press.

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7,* 52-75.

Nation, I.S.P. (2012). *The BNC/COCA word family lists* (17 September 2012). Unpublished paper. Available at: www.victoria.ac.nz/lals/about/staff/paul-nation

Nation, I.S.P., & Webb, S. (2011). *Researching and Analyzing Vocabulary*. Boston, MA: Heinle.

Nekrasova, T. (2009) English L1 and L2 speakers' knowledge of lexical bundles. *Language Learning, 59,* 647-86.

Neufeld, S., & Billuroğlu, A. (2005). *In search of the critical lexical mass: How 'general' is the GSL? How 'academic' is the AWL?* Available at http://s3.amazonaws.com/academia.edu.documents/2951027/8985atj340yff5z.pdf?AWSAccessKeyId=AKIAJ56TQJRTWSMTNPEA&Expires=1471731445&Signature=GnJQ9zBXfBmTYDbXmsU9PiFUO7w%3D&response-content-disposition=inline%3B%20filename%3DIn_search_of_the_critical_lexical_mass_H.pdf

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences*, *12*, 237–

241.

Ortega, L. (2011). *New trends in SLA research: Theories, methods, ethics*. Invited lecture at National Tsing Hua University, Taiwan, June 8.

Parrot, M. (2000). *Grammar for English Language Teachers*. Cambridge: Cambridge University Press.

Pawley, A. and F. Syder (1983) Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards and R. Schmidt (eds.), *Language and Communication*. London: Longman. 191-226.

Peters, A. (1983) *The Units of Language Acquisition*. Cambridge: Cambridge University Press.

Pinget, A-F., Bosker, H-R., Quene, H., & de Jong, N. (2014). *Language Testing, 31,* 349-365.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effects sizes in L2 research. *Language Learning, 64*, 878-912.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*, 27-57.

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics in Language Teaching, 43,* 1-32.

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2015). Lexical profiles of comprehensible second language speech. *Studies in Second Language Acquisition*, 1-25.

Saito, K., Trofimovich, P., & Isaacs, T. (2016a). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at

different ability levels. *Applied Psycholinguistics, 37*, 217-240.

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016b). Lexical correlates of comprehensibility versus accentedness in second language speech. *Bilingualism: Language and Cognition, 19*, 597-609.

Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research, 27*, 343–360.

Schmitt, N. (2010) *Researching Vocabulary: A Vocabulary Research Manual*. New York: Palgrave Macmillan.

Schmitt, N. and R. Carter (2004) Formulaic sequences in action: An introduction. In N. Schmitt (ed.), *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins. 1-22.

Selinker, L. (1972). Interlanguage. *IRAL, 10*, 209-241.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing. *TESOL Quarterly, 27*, 657-677.

Simpson-Vlach, R. and N. Ellis (2010) An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31,* 487-512.

Svartvik, J. (1980). Well in conversation. In S. Greenbaum, G. Leech & J. Svartvik (Eds.). *Studies in English Linguistics for Randolph Quirk* (pp. 167-177). London: Longman.

Stubbs, M. (2007). On texts, corpora and models of language. In M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert (Eds.) *Text, discourse and corpora*. New York: Continuum.

Stubbs, M. (1993). British traditions in text analysis: from Firth to Sinclair. In M. Baker, F Francis & E. Tognini-Bonelli (Eds.), *Text and Technology: In Honour of John Sinclair* (pp. 1-46). Amsterdam: John Benjamins.

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

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition, 15,* 905-916.

Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it. In N. Schmitt (Ed.) *Formulaic sequences: Acquisition, processing, and use* pp. *9*, 153.

Wen, Q., Wang, L., & Liang, M. (2005). Spoken and written English corpus of Chinese learners. *Foreign Language Teaching and Research Press:* Beijing.

Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary. *Behavioural Research Methods, Instruments and Computers, 20*, 6-11.

Wood, D., & Appel, R. (2014). Multiword constructions in first year business and engineering university textbooks and EAP textbooks. *Journal of English for Academic Purposes, 15,* 1-13.

Yao, Z., Saito, K., Trofimovich, P., & Isaacs, T. (2013). Z-Lab [Computer software]. Retrieved from https://github.com/ZeshanYao/Z-Lab

Yeung, L. (2009). Use and misuse of 'besides': A corpus study comparing native speakers' and learners' English. *System, 37*, 330-342.

Appendices

Appendix A: Sample target sequences, with corresponding corpus-based statistics

Sequence	Frequency	MI	FTP	BTP
the extent to which	1746.0	14.76	1.00	0.98
it would have been	1382.0	15.44	0.52	0.20
there is no doubt	721.0	19.53	0.06	0.99
in the wake of	729.0	12.79	0.99	0.99
in the house of	898.0	8.79	0.29	0.22
it may be that	758.0	12.78	0.14	0.89
with the help of	844.0	10.72	0.99	0.72
in the sense that	830.0	11.36	0.57	0.91
but it is not	790.0	11.77	0.13	0.08
on the other hand	5284.0	17.51	0.71	0.98

Note. MI = mutual information, FTP = forward transitional probability, BTP = backward transitional probability.

Appendix B. Identified Linking Adverbials

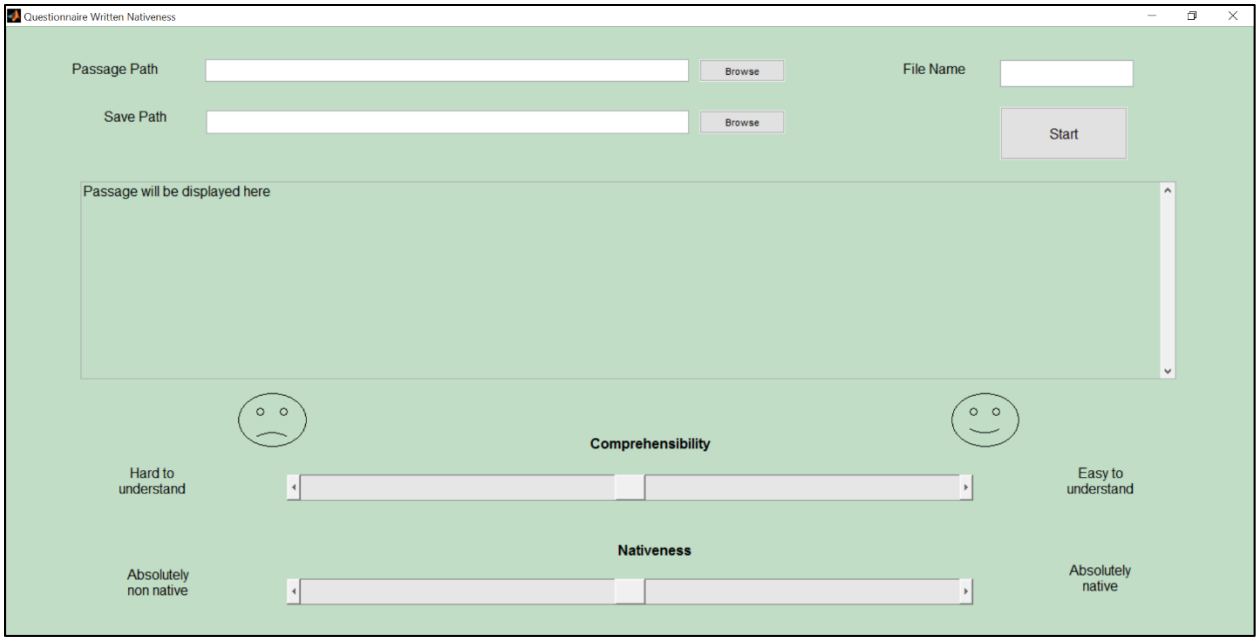
Above all
Actually
Add to this
Additional to
Additionally
Admittedly
After then
Again
All in all
Also
Although
Altogether
And
As a consequence
As a final problem
As a matter of fact
As a result
As a second drawback
As we know
As well
At the same time
At this time
Besides
But
By contrast
By doing so
By doing this
Clearly
Consequently
Contrarily
Conversely
Despite that
Especially
Finally
First
First of all
Firstly
For a start
For another
For example
For instance
For one thing
for this
For this reason
Fortunately
Further
Furthermore
Generally
Hence
However
In addition
In brief
In comparison
In conclusion
In contrary
In contrast
In fact
In general
In most case
In other words
In real life
In reality
In short
In summary
In that case
In that way
In the first place
In the same time
In the same way
In the second place
In this way
Indeed
Instead
It is true
Last
Last but not least
Lastly
Lately
Likewise
Logically
Meanwhile
More explicitly
More importantly
More precisely
More specially
Moreover
Namely
Nevertheless
Nonetheless
Obviously
On one hand
On the contrary
On the other hand
On top of that
Otherwise
Overall
Rather
Recently
Second
Second of all
Secondly
So
Stating that
That is to say
The fact is
The first
Then
Therefore
Third
Thirdly
Thus
To begin
To conclude
To do so
To finish with
To sum up
To the contrary
Undoubtedly
Unfortunately
What's more
Yet

Appendix C. Functionally categorized linking adverbials

Functional Category	Linking Adverbial	Arabic	Chinese	French
Listing				
Additive	also	20.3	6.6	7.4
	and	8.0	1.3	4.7
	further(more)	10.9	9.2	7.9
	in addition	16.7	7.9	6.7
	moreover	18.1	17.1	20.7
Enumerative	firstly	0.7	2.6	4.7
	finally	5.1	3.9	12.7
	first (of all)	5.1	11.8	18.7
	second (of all)	5.1	3.9	4.0
	secondly	1.5	3.9	9.4
Summative	in conclusion	16.7	12.5	4.7
	to conclude	4.4	2.0	11.4
Appositional	for example	23.2	18.4	23.4
	for instance	8.0	5.9	16.1
	in fact	13.8	8.5	32.1
	in other words	2.2	2.6	5.4
	indeed	5.1	5.3	22.8
Resultative	as a result	9.4	9.9	8.0
	consequently	8.7	6.6	4.0
	hence	2.2	4.6	7.4
	so	6.5	6.6	4.7
	therefore	37.7	34.8	29.5
	thus	10.9	19.7	21.4
Contrastive	actually	5.8	2.6	6.0
	but	2.2	5.3	5.4
	however	47.2	67.0	50.2
	nevertheless	3.6	12.5	8.0
	on the other hand	12.3	4.6	5.4
	yet	3.6	0.7	2.7
Transitional	besides	1.5	3.3	2.0
Total		316.4	301.6	367.5

Appendix D: Sample Training Materials and On-screen Rating Interface

Word	Explanation
Comprehensibility	Refers to ease/difficulty of understanding. If the sample is easy to understand, then the speaker is highly comprehensible. If you struggled to understand and needed to frequently re-read sections, or simply could not understand what was being said, then the speaker has low comprehensibility. 1 = *hard to understand*, 1000 = *easy to understand*
Nativeness	Refers to how closely the language resembles that of a native speaker (a user whose first language is English). If the sample reads like someone who has spoken English since birth, it would be rated as highly nativelike. If the sample reads like that of a new speaker of English, it would be rated as having low nativeness. 1 = *absolutely non-native*, 1000 = *absolutely native*



Appendix E: Full Correlations Between Lexical Variables for Picture Description and TOEFL Integrated Tasks

Picture Description Task

Lexical variable	Comprehensibility	Nativeness
Comprehensibility	—	.93**
Nativeness	.93**	—
Word length	.31**	.30**
MTLD	.31**	.32**
Causal connectives	.06	.06
Logical connectives	.11	.10
Additive connectives	–.05	–.11
Causal verbs	.00	.01

Causal verbs & particles	.08	.07
Verb overlap	–.27**	–.24*
Modifiers per noun phrase	–.13	–.14
Gerunds	–.26*	–.19
Infinitives	.22*	.27**
Word frequency	–.28**	–.25*
Word age of acquisition	–.22*	–.20
Word familiarity	–.10	–.11
Word concreteness	–.24*	–.21*
Word imageability	.19	.18
Word meaningfulness	.23*	.19
NAWL	.24*	.28**
NGSL_1	–.16	–.12
NGSL_2	.08	.03
NGSL_3	.08	.10
NGSL_Off	.05	.01
BNL_0	–.23*	–.18
BNL_1	.05	.04
BNL_2	–.04	–.07
BNL_Off	–.02	–.04
BNC/COCA_1	–.09	–.05
BNC/COCA_2	.11	.04

BNC/COCA_3	.06	.08
BNC/COCA_Off	.13	.11

Note. *p <.05, **p < .05, two-tailed.

TOEFL Integrated Task

Lexical variable	Comprehensibility	Nativeness
Comprehensibility	—	.94**
Nativeness	.94**	—
Word length	.29**	.33**
MTLD	.49**	.44**
Causal connectives	.18	.12
Logical connectives	.05	.01
Additive connectives	–.25*	–.21*
Causal verbs	.17	.22*
Causal verbs & particles	.23*	.18
Verb overlap	.08	.04
Modifiers per noun phrase	–.25*	–.22*
Gerunds	.05	.06
Infinitives	.03	.02
Word frequency	–.22*	–.20
Word age of acquisition	.27**	.27**
Word familiarity	–.27**	–.30**

Word concreteness	–.14	–.11
Word imageability	–.14	–.13
Word meaningfulness	–.09	–.10
NAWL	.07	.07
NGSL_1	–.11	–.12
NGSL_2	.18	.19
NGSL_3	.31**	.34**
NGSL_Off	–.23*	–.24*
BNL_0	.03	.07
BNL_1	–.17	–.21*
BNL_2	.23*	.27**
BNL_Off	–.33**	–.35**
BNC/COCA_1	–.16	–.17
BNC/COCA_2	.21*	.23*
BNC/COCA_3	.26**	.28**
BNC/COCA_Off	–.34**	–.36**
Lecture overlap	–.21*	–.22*
Reading overlap	–.02	–.05

Note. *p <.05, **p < .05, two-tailed.