Brigham Young University
BYU ScholarsArchive

All Theses and Dissertations

2008-11-25

The Effects of Homography on Computer-generated High Frequency Word Lists

Athelia Graham
Brigham Young University - Provo

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

 Part of the Linguistics Commons

BYU ScholarsArchive Citation

Graham, Athelia, "The Effects of Homography on Computer-generated High Frequency Word Lists" (2008). *All Theses and Dissertations*. 1617.
https://scholarsarchive.byu.edu/etd/1617

This Thesis is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu.

THE EFFECTS OF HOMOGRAPHY ON COMPUTER-GENERATED HIGH FREQUENCY WORD LISTS

by

Athelia Graham

A thesis submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Arts

Department of Linguistics and English Language

Brigham Young University

December 2008

Copyright © 2008 Athelia Graham

All Rights Reserved

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Athelia Graham

This thesis has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

_____	_____
Date	Dee Gardner, Chair
_____	_____
Date	Mark Davies
_____	_____
Date	Neil J. Anderson

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Athelia Graham in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

_____ _____

Date

Dee Gardner
Chair, Graduate Committee

Accepted for the Department

_____

William Eggington
Department Chair

Accepted for the College

_____

Joseph Parry
Associate Dean, College of Humanities

ABSTRACT

THE EFFECTS OF HOMOGRAPHY ON COMPUTER-GENERATED HIGH FREQUENCY WORD LISTS

Athelia Graham

Department of Linguistics and English Language

Master of Arts

This study investigated the significance of semantics in computer-generated word frequency counts in response to a call for new word lists (Read, 2000; Gardner, 2007). Read claims that no corpus projects to date have produced any "definitive, stand-alone word-frequency lists" (p. 226). Many researchers are wary of the fact that the concept of a word is never clearly defined in most studies that have dealt with word frequency counts. It is clear from the research that one universally acceptable construct for the concept of *word* does not exist. In fact, many past word frequency counts only examine word forms without considering the word meanings and the possible effects of homography on lists.

Ming-Tzu and Nation (2004) did some research on the Academic Word List (AWL) that addresses some criticisms of word-frequency lists. They evaluate the extent of homography throughout the AWL. However, words found in the AWL are often not a part of the highest frequency word-forms in English.

The present study focuses on high frequency words. It evaluates a randomized sample of 46 lemmas that occur at least 1500 times in the British National Corpus (BNC).

A further random sampling of 200 examples for each lemma, in context, was semantically analyzed and tallied. One hundred of these examples were from the written portion and the other 100 from the spoken portion. The list of meanings for each word was compiled using conflated WordNet senses and some additional senses. Each context was double and sometimes triple rated. The results indicate that the impact of semantic frequency versus form-based frequency is considerable. The study suggests that the presence of homography tends to be extensive in many high-frequency word forms, across major registers of the language, and within each of the four major parts of speech. It further suggests that basing frequency on semantics will considerably alter the content of a high-frequency word list.

ACKNOWLEDGEMENTS

My gratitude is extended to many people for their help with the research and patience with me on this project. First, I would like to thank Dr. Gardner for his patience through this entire process, for his steady and persistent encouragement, for his dedication to quality research, for his contagious passion for vocabulary, for research, and for learning. I would also like to thank Dr. Davies for all of the expertise and help he provided in the area of corpus linguistics and in preparing the data for analysis and Dr. Anderson for his encouragement and optimism over the years of completing this as well as for the opportunity to apply what I have learned from my thesis through the vocabulary project being done at the ELC. Additionally, I would like to express gratitude for all of my professors and classmates who I have been able to learn from and grow with and feel that I have study among amazing people. I would also like to extend thanks to all of my friends and colleagues who were willing to help with the ratings and revisions. Finally, I would like to express special thanks to my parents for their unconditional love and support, and their many hours of service during this process: their continual advice along the way, the hours of conflating, rating, and analyzing data, and their extremely valuable help with editing and revising during the writing process.

Table of Contents

List of Tables

List of Figures

CHAPTER ONE

Introduction

In recent years, many studies have shown the essential role that vocabulary plays in both first language (L1) and second language (L2) acquisition at all proficiency levels. Vocabulary is a major element that allows speakers to relay meaning and ideas. In a first language, people's depth and breadth of vocabulary knowledge have proven to be one of the most important indicators of their intelligence level and achievement (Laufer, 2003; Nagy & Herman, 1987; Vermeer, 2001). For L2 learners, vocabulary is integral to communication and advancement in the target language (TL). For example, without any knowledge of grammar, syntax, or morphology, beginning language learners can say "restroom" or "toilet" and communicate enough to get them to the place they need. For advanced learners, the ability to express ideas, read with more understanding, and write with more fluency and more advanced thought processes in an L2 are greatly facilitated or hindered by their L2 vocabulary knowledge.

In everyday exposure to language, a large vocabulary is necessary in order for people to carry out general tasks such as reading an article in a newspaper or magazine, glancing through a website on the internet, talking to someone on the phone, listening to a lecture or a program on the radio, or writing a research paper. All of these tasks will expose people to a wide range of vocabulary items that they must be familiar with in order to comprehend and communicate effectively, though even native speakers sometimes do not recognize, do not understand, or do not have an exact knowledge of some words they encounter.

For ESL learners, a long-range goal of approaching the vocabulary size of a native speaking adult creates an enormous learning burden, especially in light of the recognized vocabulary deficiency in areas such as L2 reading proficiency (e.g. Nation, 2006; Pulido, 2003). Nation's (2006) research has indicated that in order for comprehension to occur in written or spoken contexts, 95–98% of the words must be known. He suggests that 95–98% coverage at a high school level translates into needing to know 8,000–9,000 word families for written texts. This most current estimate of vocabulary items ESL students must learn in order to have an acceptable level of competency have slowly been rising from the original estimates of 2,000 – 3,000 word families due to various characteristics of vocabulary that have gone unaccounted for in computer-generated vocabulary studies and word lists.

In order to learn about word usage in authentic texts, linguists have developed computer-based methods of examining language. They have been able to collect large bodies of language (mega-corpora), and analyze them using sophisticated programs to produce frequency counts of word usage in both spoken and written texts.

In particular, the use of corpora has brought to light various trends in the frequency counts. For example, certain words occur with great frequency in a given text while others may occur only once or not at all. Since studies have also shown that the frequency of occurrence of words affects the likelihood of them being acquired (Laufer, Elder, Hill, & Congdon, 2004 and Read, 1988 as cited by Nation, 2006), the obvious strategy for teachers and learners is to deal first with the words that occur most frequently. Recent computer-generated frequency counts from large mega-corpora have

been used regularly by educators wanting to make more informed vocabulary selections for teaching in their ESL programs.

Despite their usefulness in the past, some major issues have surfaced with regard to the validity of electronically-based word frequency counts and the subsequent creation of high frequency word lists. These issues have caused researchers to come up with widely varying numbers and results concerning how many words native speakers know and how many words ESL learners need to know for adequate comprehension and specific levels of proficiency (see Nagy and Anderson, 1984; Nation, 2001a, 2006).

Perhaps the most wide-ranging problem among the various counts and lists is the lack of agreement about what to count as a word, lexical item, or vocabulary unit. This may seem simple to answer, but there are many questions to consider both with regard to learners' perceptions and to the connection between word forms and their meanings.

The following questions and examples illustrate these issues. Should words be counted as individual forms, as lemmas, or in word families? Should phrases like *by and large* or phrasal verbs like *crack down* be counted as a whole or as individual parts? If ESL learners know the word *father* (a male parent), can it be assumed that they also know *fathers* (the plural form), *father* (an ecclesiastical leader), and *godfather*? The semantic relationship between some occurrences of a word is very transparent, while it is much less obvious for others. In the case of *bat* (the animal), *bat* (in baseball), and *bat* (to flutter the eyelashes) is a new word learned with each distinct meaning related to that form? Do *father*, *fathers*, and *to father a child* all count as one word or as separate words? If learners know the color *blue*, how transparent is the meaning of *blue* in the phrase *he was feeling blue today*? Will learners understand the meaning of *kick the*

bucket (to die) because they know the individual meanings of *kick*, *the*, and *bucket*? Will they know *crack up* (to laugh) because they know what *crack* and *up* mean? Can learners make a connection between *make*, *make up* (to invent), *make up* (to reconcile), and *makeup* (cosmetics)? Of the many issues implied in these examples, only two will be addressed here due to the limited scope of this study: 1) how should semantic boundaries of distinct words be determined?, and 2) what concept of a word should be used as the unit of measurement in a frequency count?

Even within these two issues, the present study will focus specifically on the existence of written word forms with distinct multiple meanings (homography), particularly in high frequency items, the extent to which it exists, and how the existence of homographs would alter the content, nature, and size of word lists used for pedagogical purposes in language learning contexts. In addition, it will look at whether the frequency of homographs differs in written and spoken language. Finally, it will assess the implications of existing homography in English texts and discourse for estimates of word coverage and the increased burden of learning and teaching high frequency vocabulary items.

Definition of terms

The following definitions will be assumed throughout the thesis:

Word: The idea of *word* refers to a form, an individual arbitrarily determined unit that is represented in specific phonological and graphic forms. There is no reference to semantics or to meanings in the definition because of the complications generated by doing so. A current and generally acceptable definition of *word* is provided by Nagy and

Anderson (1984): “a graphically distinct sequence of characters bounded right and left by a space” (p. 306).

Sense: The term *sense* will be used interchangeably with the concept of meaning.

Lexical Item: The term *lexical item* in this thesis refers to a semantic unit that may consist of one word or a group of words which represent one idea, meaning or sense.

Lemma: A *lemma* is a specific word form and its inflections, without consideration of meaning. It is represented in small capital letters followed by its part of speech throughout this study. For example, the lemma DEVELOP (v) includes the forms *develop, develops, developed, developing*. On the most basic level, lemmas are often limited to one part of speech. For example, the noun *water* would be a separate lemma from the verb *water*. In this study, any word in all small capital letters (i.e. DEVELOP (v)) will be used to represent lemmas (Knowles & Mohd Don, 2004; Stubbs, 2002; etc.).

Lexeme: A *lexeme* refers to a semantic unit and signifies a lemmatized form with a distinct meaning. It is represented by a lemma in brackets []. For example, the lemma FAIR (v) can be separated into several lexemes: [FAIR] (adj) = just; [FAIR] (adj) = beautiful; [FAIR] (adj) = light skinned; [FAIR] (adj) = mediocre; [FAIR] (adj) = reasonable or not extreme; [FAIR] (adj) = no clouds in the sky; [FAIR] (adj) = legitimate hit in baseball.

Word Family: A *word family* is the most liberal grouping of word forms because it primarily focuses on morphologically related forms, without separating parts of speech. Bauer and Nation (1993) define a word family as “a base word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately” (p. 253). The basic idea here is that there is a core or central form and meaning from which certain derived forms with their meanings are transparently

connected and therefore should be included in that word family. For example, the word family for *develop* in Bauer and Nation's study includes the following forms: *develop, develops, developed, developing, (un)developable, developer(s), undeveloped, development(s), developmental, developmentally, developmentwise, semideveloped, antidevelopment, redevelop,* and *predevelopment* (p. 254). However, with regard to electronically realized words with no semantic tags, the definition of word family can have no reference to meaning, and by implication, can only be bare forms plus their inflected and derived forms. Thus, a word family generated from an electronically-compiled data base would not distinguish between lexemes like [DEVELOP] (v) = to produce pictures from film through a chemical process, [DEVELOP] (v) = to make land available and useable, and [DEVELOP] (v) = to create a strategy or theory, etc.

Polysemy: In this study, *polysemy* will be defined as the concept of a word form having several related meanings. This term includes the entire continuum of meanings from subtle nuances distinguished because of context (e.g. *big*: tomorrow is a *big* day for him; winning gave him a *big* head; he is a *big* spender), to perhaps more distantly related meanings of word forms in which semantic connections may still be somewhat transparent (e.g. *hand*: a good *hand* in cards; give me a *hand* with this heavy couch; he shook my *hand;* let's give her a *hand* (applause) for a job well done).

Homography: *Homography* as used here applies to words that have the same written form but have separate and distinct meanings. For instance, the word form *bat* is used to show four distinct *homographs*:

1. He *bats* the ball well.
2. They sell three different brands of baseball *bats*.
3. The *bats* flew out of the cave.
4. She flirtatiously *bats* her eyelids.

The first two examples are within the same semantic boundaries but in different word classes. Examples two and three fall in the same word class, but are semantically unrelated. Example four is from the same word class as example one, but has no transparent semantic relationship with any of the other examples.

High Frequency Words: For this study it is important to establish what is meant by the term *high frequency* words. Some researchers in the past have looked at this idea with a minimalist approach, suggesting a vocabulary of the most frequent and productive words to define this idea of *high frequency*. For example Ogden (1934) suggested that 850 words were all that was needed to communicate effectively in English. West (1953) claimed that around 2,000 words essentially made up the core of English and provided the necessary jump start into communication for learners of the language. However, more recently researchers have begun to look at what is termed the coverage of vocabulary. That is, they look at how many words cover a certain percentage of written or spoken texts.

In one study, Coniam (1999) refers to frequency bands and sets the cut-off point for high frequency words in his study at the band of 80% coverage. On his chart of frequency bands this includes lemmas defined as extremely frequent, very frequent and frequent (p. 106). This means that a high frequency word list would include all of the lemmas from a raw lemmatized frequency list that it took to cover 80% of a text, or other representative corpus of the English language. Using frequency counts from the Bank of English corpus, he determined that 80% text coverage was possible with 2,145 lemmas.

Nation (2001a) suggests five other possible ways to make a distinction between high and low frequency words. He points out that “frequency studies show that there is

no clear dividing line between high and low frequency words," meaning that the dividing point is made somewhat arbitrarily (p. 168). He suggests the importance of clearly defining a high frequency vocabulary, especially for those who may only have limited time and exposure to English. He points out that the value of a valid and reliable high frequency word list is that students who focus on those words can get more use from these highly productive words for the effort they put in to learning them.

Nation's (2001a) first suggestion for determining high frequency words is to create a list of words that represent 95% coverage of a representative English text. Some of the problems he finds with this approach are that the number varies depending on what the target texts are. If a corpus with adult language and/or academic language is used, the number of words needed for 95% coverage increase to 14,000. The second suggestion is to look at the cost (effort to teach the words) and benefit (coverage) of words in order to include them on the high frequency list. This method places primary emphasis on the percentage of coverage of each lemma in a text which averages 300 words per page. Using the Brown Corpus, he found that a satisfactory cut-off point for defining *high frequency* words would be at around 3,000 words (p. 174). This was based on the fact that on a 300 word page, any lemmas beyond 3,000 would not appear on average at least once per page. The third way Nation recommends to delimit high frequency lists is by comparing several word lists that come from linguistically differing corpora (e.g. covering different registers) and creating a composite list from the overlapping words. The weakness with this method is that the content of and purpose for creating the word lists (and for the selection of texts put into the corpora) will substantially influence what words are found on it.

Nation (2001a) proposes a fourth method of determining a high frequency word list, which is more practical from a teaching standpoint, but less valid theoretically. Selection for such a list is simply based on what teachers find is a realistic number of words to teach and learn in their program, depending on factors such as the length of the program, the level of students, and the language focus of the ESL class. The fifth and final method he suggests is to create a core vocabulary, similar to Ogden and Richard's creation of Basic English in 1943, which included only 850 words that are highly productive. However, Nation calls attention to the fact that Basic English has been highly criticized because of its lack of practicality and suggests that the idea of a core vocabulary can be as subjective and ambiguous as the idea of high frequency words.

Many high frequency lists seem to set a maximum of 2,000 words. Nation (2001a) proposes that, "it seems sensible to have a high frequency word list of 2,000 [word families]" (p. 179). This 2,000 word threshold is considered "sensible" as a standard for perhaps two primary reasons: 1) around 2,000 words have been shown to cover at least 70 – 80% of running texts in English, and 2) 2,000 words may be the maximum manageable limit of vocabulary to teach and learn over the course of an English program. Also, several of the benchmark word frequency studies of the past have commonly maintained a size of 2,000 or less, such as Thorndike and Lorge (1942) and West (1953). Nation (2001a) points out that after the first 2,000 words, it is difficult to determine frequency because words have a narrower range, thus making the text selections included in corpora influential in the words found in frequency levels following the 2,000 word level. Read (2000) agrees, making the point that "the further we move from the first 2,000 [words] or so, the less significant frequency becomes in an

absolute sense" because "the selection of lower-frequency words depends increasingly on the learners' specific needs and interests" (p. 228).

Because of the ambiguity concerning the definition of a *high frequency word (or lemma)*, the limit used in this study was set as any lemma occurring 1,500+ times in the BNC. This criterion encompasses the top 4,277 lemmas of the BNC. This is more liberal than the 2,000 word standard mentioned above. However, this limit still excludes some fairly basic words from Kilgariff's lemmatized frequency list of vocabulary from the BNC (http://www.kilgariff.co.uk/BNC_lists/lemma.al), such as *planned*, *grandfather*, *ill*, *coming*, and *closed* that intuitively seem to be common, but do not quite make it into the criterion of occurring 1,500+ times. Though the 1,500+ threshold is somewhat arbitrary, it seemed wise to take a more liberal approach than the standard 2,000 lemmas or word families for at least two reasons. First, as research advances, it seems that the number of word families or lemmas that people need to know for reading comprehension thresholds is inconsistent (e.g., Nation, 2001a and Nation, 2006), is usually underestimated, and is in dispute. This is in large part due to the fact that researchers cannot agree on what they are counting as a word (i.e. word form, meanings, multi-part items, lemmas, word families, etc.). Second, because this thesis only analyzed a small sample of words it seemed better to err on the side of using a larger grouping of words and thus avoiding the possibility of excluding possibly relevant words from the analysis.

CHAPTER TWO

Review of Literature

Introduction

As was mentioned in the previous chapter, the purpose of this thesis is to explore the effects of counting distinct word meanings on computer-generated and electronically based high frequency word lists. This chapter will begin by giving a brief history of corpora and word frequency counts in order to see what has led to the current state of affairs. Following this, the critical issue of the relationship between form and meaning will be discussed as it relates to vocabulary acquisition for second language learners, the existence of homography and polysemy in English vocabulary, and the concept of semantic relatedness. Further, there will be a discussion on how those linguistic characteristics influence the construct of *word* and how that directly affects computer-generated high frequency word counts and in particular, those intended for ESL instruction and learning. Finally, the specific questions to be addressed in this thesis will be presented.

Corpora and Word Lists – A brief history

In looking at the creation of word frequency lists of English (and particularly at how homography has been dealt with – or ignored) over the past 100 years, it is important to understand that the task of creating a "good quality list" is very complex, as Read (2000) suggests. He asserts that it requires the consideration of many possible variables, such as having a clear purpose for the list and having a solid understanding of the form-meaning relationships of words that are formed during the process of vocabulary acquisition inside a language learner's mind.

Though the field of corpus linguistics seems fairly new, the earliest corpus study of English word frequency counts mentioned in the literature was that of Kaeding in 1889, in which he used a corpus of approximately 11 million running words in "lexicometrical research" (Engels, 1968, p. 213). Such research continued throughout the rest of the 20th century. The early lists were done by hand since they pre-dated computer development and availability. One of the most influential of these early lists was done by West (1953). He used a hand-counted frequency list from Lorge (finished in 1949) to create a semantically tagged and organized list of high frequency words which he called the General Service List (GSL). He envisioned this list to be "the selection of English most suitable to set as a first objective for foreign learners" and as "trying to simplify English for the learner" (1953, p. iv-v).

West's GSL has been criticized as outdated because the corpus on which it was based was smaller than the mega-corpora of today, and was created only with written language samples of English from nearly a century ago (Coxhead, 2000; Engels, 1968). However, West seemed to have understood that semantic analysis of the words was an essential part of creating a valid high frequency word list. Even fifteen years later, Engels (1968) recognized the significance of this when he pointed out that many of the word lists were deficient because they "treated the word as an objective symbol, neglecting the distinction between semantic contents in each word" (p. 214).

This deficiency pointed out by Engels has been further perpetuated in English frequency word lists as computers have continued to develop a greater capacity to process linguistic data. Starting in the 20th century, a plethora of large corpora began to emerge: the Brown corpus in the 1960s (and the FROWN in the 1990s); the LOB in the

1970s (and the FLOB in the 1990s), which dealt with written text only; the London-Lund corpus (half written and half spoken) in the 1980s; the Australian Corpus of English (ACE), also in the 1980s; the BNC (90% written and 10% spoken) in the 1990's. With several of these corpora, and even with the dramatically larger BNC, grammatical tagging was done, using the well-known CLAWS tagger. But still, these corpora and the ensuing word frequency lists lacked semantic tagging. The same problem continues as larger and larger mega-corpora are being created, such as the Oxford English Corpus (2000-2006), which consists of over 2 billion words, and is not even grammatically tagged. Similarly, the most recently finished mega-corpora to date, the Corpus of Contemporary American English (2008), consisting of 385+ million words, has been grammatically tagged (like the BNC), but also lacks semantic tagging (http://www.americancorpus.org).

Unfortunately, as computers have become capable of crunching large amounts of linguistic data, word forms have become the priority, while semantics has taken a back seat and often been totally ignored. Nor have any computer-based programs been developed to adequately disambiguate senses in order to effectively execute semantic tagging of large corpora. This is largely due to many complexities inherent in English vocabulary, including homography, form-meaning relationships, multi-word lexical items, lexico-grammatical relationships, and the relationship of word meanings to their surrounding contexts. Furthermore, few attempts have been made to semantically tag corpora by hand – most likely because it is an extremely overwhelming and time-consuming task. Thus, in spite of advancements in the creation of more up-to-date corpora and word frequency lists created from them, the lack of this key connection

between form and meaning has persisted and skews the content of word frequency lists and their validity as ESL language teaching and learning tools.

Some Problems with Frequency Counts

As potentially helpful as word frequency lists can be in determining which words are most important for language learners to acquire, there are various inherent problems with creating and using them that affect the usefulness of such word lists. One of the main problems with frequency counts is that frequency is directly determined by the language samples chosen to be in the corpus to represent the language (Harris & Jacobsen, 1974; Nation, 2001a; Stubbs, 2002). This brings to light various issues, such as datedness of sources, register representativeness, and target linguistic populations. For the creation of any frequency list, the important question to ask is "What English is this representing?" Even a frequency count that is based on a claimed general body of language is subject to bias because of the selections made by the individuals creating the corpus and their reasoning behind those selections. All frequency lists have this limitation.

For example, the BNC is meant to be a representative corpus of general English. Yet, the contents included in the corpus (academic journal articles, newspapers, magazines, novels, etc.) clearly contain vocabulary more familiar to a well-educated adult population. It has a shortage of lexical items learned by most adults in their school-aged years, which are considered general and common, and which all adults are surely expected to know, but which may not show up frequently in a corpus centered on educated adult language. This problem becomes even more convoluted as homographs are considered because all of the distinct senses of a word-form are counted as the same

word (*my mother was fair* – meaning either *just* or *had light skin*). Thus, frequency counts may often under-represent some words and overlook many others.

One other issue to consider with regard to research and word frequency counts is that the concept of frequency has been operationalized in many different ways. For example, in looking at how frequency affects the learnability of a word, some researchers have looked at frequency based on L1, some on L2. Others have used raw form frequency, while others have counted lemmas, or word families. This presents an obstacle that manifests itself in the studies that use frequency as an independent variable. With word lists, the primary complication of measuring frequency lies in the variety of ways researchers have chosen to define the construct of a word (Gardner, 2007; Read, 2000). This issue is directly affected by the inherent complexity of the relationship in all languages between form and meaning.

The Problem of Form and Meaning: Implications for ESL Learners

Many linguists have recognized the complexities of form-meaning relationships in vocabulary, particularly in second language acquisition (e.g. Gardner, 2007; Nation, 2001b; Nerlich, Todd, Herman, & Clarke, 2003; Read, 2000; Sinclair, 2004; Stubbs, 2002; Zoughoul, 1991). Because the main rationale for carrying out this study is to evaluate the quality of current word lists and improve the process of creating word lists for ESL teaching and learning, it is also important to understand the challenges of the form-meaning relationship on a psychological level for the learner as well as on a theoretical linguistic level.

First, it is important to briefly describe the involvedness of ‘knowing’ a word. Knowing a word implies many things. Miller (1999) cites five types of word knowledge:

"the ability to define it, the ability to recognize situations for using it, knowledge of its alternative meanings, the ability to recognize inappropriate uses of the word, and the availability of the word for use in everyday life" (p. 2). Haastrup and Henricksen (2000) suggest another way of measuring people's word knowledge by looking at three different dimensions of lexical competence: "(1) partial – precise (different levels of comprehension of the same lexical item), (2) receptive – productive, (3) depth of knowledge" (p. 222). There are obviously numerous factors involved in knowing a word. As is suggested, there are various dimensions and aspects of lexical development and thus the concept of 'knowing' rests on a continuum of proficiency for each word form that exists. As Haastrup and Henriksen point out, learning a word is not a linear process. This is an important factor to understand with regard to the psychological realities of vocabulary acquisition in an L2 and the increase in psychological demands that homonymy places on that acquisition process.

Often, the native-speaking teachers are looking through a lens of high linguistic awareness and familiarity. Thus, when approaching the task of teaching L2 learners, the perspective of the native speaker is skewed with regard to the psychological realities and difficulties of learning vocabulary and making conceptual connections with appropriate word forms. This constitutes one of the great challenges of linguists and others who have a hyper awareness of the English vocabulary and its existing form-concept network of relationships. As lexical boundaries are defined, the psychological realities of form-concept relationships for a native speaking linguist are potentially very different than the psychological realities of learning linguistically appropriate form-concept relationships for L2 learners (Gardner, 2007).

Understanding the extent to which an L1 affects L2 vocabulary acquisition is important in making decisions about distinguishing homographs, specifically with words that have a large spectrum of meanings that could be somewhat related or completely unrelated, such as the word *bear* and all of its word forms, which will be demonstrated in the next section. Nation (2001b) suggests that "the more a word represents patterns and knowledge that learners are already familiar with, the lighter its learning burden" (p. 23-24). Later in the chapter Nation (2001b) continues with this concept of the 'learning burden' of a word, saying that "making the form-meaning connection is easier if roughly the same form in the first language relates to roughly the same meaning" (p. 48), making loan words and cognates a lighter learning load because the form-meaning relationship is already established in the mental lexicon. Some studies go further in explaining that differences in L1 and L2 semantic boundaries and sense-form relationships, emphasizing the fact that the conceptualization of ideas in an L1 indeed have a direct effect on lexical development in an L2 (e.g. Ijaz, 1986; Zughoul, 1991). Thus, both defining semantic distinctions between homographs, such as *bear* the animal and *bear* the verb, as well as making semantic connections between polysemes, such as *bore a burden*, *bear with me*, and *the child was born*, can affect the psychological realities of linguistic knowledge, connectedness, and accessibility for L2 learners, and can vary considerably depending on their native language.

A study done by Zughoul (1991) exhibited written errors in lexical choices by Arabic speaking ESL students. He drew some interesting conclusions about the form-meaning relationship and semantic boundaries and how they both can cause problems for ESL students. One error he found to be quite common was confusion with words that

had similar forms. He found that the L2 learners were selecting the wrong lexical item due to phonetic and graphic similarities between the two forms. For instance, one student wrote "People are unable to work and earn efficient money" (p. 52), meaning *sufficient* instead of *efficient*. This specific example shows that form alone, without taking into account semantic aspects, can increase the learning burdens of words.

Another example in Zughoul's (1991) study reiterates the difficulty of understanding semantic boundaries at a productive level. Of the top ten most common lexical errors he found in students' writing, the most common error, by a large margin, was "assumed synonymy" meaning that the students "assume that a number of related words are synonymous to the extent where they can be used interchangeably" (p. 48). He indicates that semantic and syntactic boundaries which exist for each lexical item are often very subtle and slight, thus making it difficult for an ESL learner to realize the distinctions between the two choices and thus make an error in lexical choice. For example, one student wrote "There are many works in the city" (p. 48) using *works* instead of *jobs*. This problem is further exacerbated by polysemous meanings of a word form that may be distinct words in an L2 learner's L1, while it is considered only a nuance of a 'core' meaning tied to a word form, or vice versa. These results and examples imply that homonymy (both phonological and graphic similarities) complicates the form-meaning relationships in English and increases the learning burden of homographs.

With regard to the specific issue of homographs and their effects on learning burden for ESL students, not many studies have been done. However, several researchers have done in-depth studies looking at the semantic values of a word, a few words, or

some morphemes (i.e., Nerlich *et al.*, 2003; Ravin and Leacock, 2000; Stubbs, 2002). However, no studies have attempted to evaluate the extent of the existence of homography on a large scale among all high frequency words. Despite the lack of research in this area, one can hypothesize that in light of the complexities discussed above concerning vocabulary acquisition for L2 learners, ESL students' native language background has a substantial impact on vocabulary acquisition. This is primarily because semantic boundaries differ from one language to the next. Thus, any additional distinct or even related meanings that must be connected to a form with an already existing semantic network is bound to increase the learning burden of that word.

The Problem of Form and Meaning: Polysemy, Semantic Relatedness, and Context

Polysemy is an important issue with regard to high frequency word counts primarily for two reasons: 1) it is directly connected with homonymy and so contributes to the increased learning burden of words for ESL learners, and 2) it complicates the process of defining the construct of *word* and consequently how words are counted and what words are included in word lists. Ravin and Leacock (2000) have done extensive research on the nature and existence of polysemy in English and suggest that "the most commonly used words tend to be the most polysemous," thus highlighting the significance this has in high frequency word counts (p. 1). In their book they make an important distinction between polysemy and homography, pointing out that polysemes can, in a manner of speaking, develop into homographs over time as their semantic relationship deteriorates:

> Strictly speaking, homographs are etymologically unrelated words that happen to be represented by the same string of letters in a

> language . . . Conversely, polysemes are etymologically and therefore semantically related, and typically originate from metaphorical usage . . . The distinction is not always straightforward, especially since words that are etymologically related can, over time, drift so far apart that the original semantic relation is no longer recognizable. (p. 2)

Thus, polysemy and homography rest on a continuum of semantic relatedness and an exact distinction between the two concepts is somewhat hazy and ambiguous. (Nerlich, *et al.*, 2003; Ravin & Leacock, 2000). And the fundamental nature of this changes even more in the context of the psychological linguistic realities of vocabulary for ESL learners.

A good illustration of this is the word *blue*. *Blue* is an adjective that can represent both the idea of a color and of the feeling of being gloomy, dispirited, or mildly depressed. For most native English speakers, the connection between the color and the feeling is so close that it seems very transparent and could likely be deemed polysemous. However, that semantic connection may be strongly bound to cultural background and psychology, thus not making sense or not seeming obvious to some ESL students who have not ever conceived of that connection. Thus, from the perspective of second language acquisition, it may be more psychologically valid to count these two meanings of *blue* as homographs for ESL instructional and learning purposes.

The relationship between polysemy and homonymy becomes important in the process of determining how to count lexical items. Ming-Tzu and Nation (2004) point out that “Polysemy and homography are points on a scale and there can be considerable

disagreement about whether two items are polysemes or homographs" (p. 295). This is because, as illustrated with the example of *blue,* the extent of semantic transparency and overlap between meanings is debatable as well.

Part of this problem is due to the inherent nature of language. The lexicon is a network of infinite concepts that exist and are somewhat sloppily assigned to a finite group of interconnected forms which vary extensively from language to language (Nerlich *et al.*, 2003). Many linguists who have studied various lexical items or characteristics of vocabulary in depth have come to recognize why the form-meaning relationship is quite difficult to standardize and define in a concrete way (e.g. Anderson & Nagy, 1991; Nerlich, *et al*., 2003; Ravin & Leacock, 2000; Sinclair, 2004; Stubbs, 2002; Wei & Light, 1973). Stubbs (2002) articulates this saying that "the central problem in linguistic description is how to describe a system which is both highly complex and highly variable" (p. 97).

Of the in-depth studies done regarding the semantics of English vocabulary, several parallel conclusions have been drawn about why the form-meaning relationship is so blurry and causes ambiguity in the distinction between homographs and polysemes. Some of the most prominent conclusions are as follows: (1) the existence of a continuum of meanings (polysemy) and multiple distinct meanings (homonymy) (Carter, 1998; Knowles & Mohd Don, 2004; Ming-Tzu & Nation, 2004; Nerlich *et al.*, 2003; Stubbs, 2002), (2) the reciprocal influence of context and individual words on each other's meanings (Anderson and Nagy, 1991; Ravin & Leacock, 2000; Sinclair, 2004; Stubbs, 2002), (3) the tendency of words to form tightly bound multi-word lexical combinations and idiomatic chunks (Anderson & Nagy, 1991; Darwin & Gray, 1999; Ogden, 1942;

Sinclair, 2004), (4) the existence of the inseparable lexico-syntactic relationship of word form, word meaning, and the linguistic (both cultural and grammatical) boundaries of lexical items (Anderson & Nagy, 1991; Sinclair, 2004; Stubbs, 2002), and (5) synonymy and metonymy (using a part of an object or idea to represent a whole; e.g. *crown* = a monarch) (Nerlich *et al.*, 2003; Stubbs, 2002).

With regard to individual lexical items, the continuum of possible meanings can be large or small, therefore making it more difficult to delineate homographs and polysemes. Extensive research in polysemy (Stubbs, 2002) and the impact of context on words (Sinclair, 2004), suggests that meanings of words are ultimately determined by the context in which they are found. Stubbs (2002) suggests that "meaning is use", explaining that "the meanings of words and phrases differs according to their use in different linguistic and social contexts" (p. 20). Thus high frequency words, because they are found in so many contexts have the potential for more polysemy and homonymy. Nerlich *et al.* (2003) point out that the "multiplication of meaning" is caused by "people's perception of meaning and then the subsequent use of a term that may alter the meaning somewhat, [and] can, over a period of time, drastically change the meaning of a word" (p. 61). This trend is perhaps more common in spoken language and especially with slang terms, such as with words like *wicked, sick,* and *tight* that currently mean *cool* or *awesome* and with expressions like *she goes* and *he's like* that now mean *she/he says*.

Another aspect of context affecting word meaning is the relationship between polysemy, synonymy, and the lexico-syntactic relationship of words. Oftentimes words have multiple related meanings that are undoubtedly influenced and often determined by their context. Anderson and Nagy (1991) illustrate this with an example using the word

give. They suggest that even though the word *give* is a synonym of *grant* and *donate*, "you can *give* someone a shove, but not *grant* someone a shove; you can *give* a performance, but not *donate* one, at least not in the same sense," and "*donate* unlike the related verb *give*, cannot take an indirect object" (p. 701). These types of synonymous relationships that are distinguished by meaning nuances and syntactic limitations are one more characteristic that complicates the form-meaning relationship of vocabulary.

Nerlich *et al.* (2003) go on to suggest that "Polysemy is pervasive in language" and "it is not just an accident of history and synchrony, but rather an essential manifestation of the flexibility, adaptability, and richness in meaning potential that lie at the very heart of what a language is and what it is for" (p. 80). Due to the fact that such linguistic characteristics and phenomena exist consistently and occur on a regular basis, defining words, determining semantic relatedness and lexical boundaries, and distinguishing between polysemes and homonyms becomes a complicated and seemingly impossible task.

Obviously, these characteristics of language complicate the efforts of lexicographers to define, categorize, and group words. One attempt at remedying this has been to try to establish a core or basic meaning that inherently exists in all of the derivations and inflections of a certain root word or base form (Anderson & Nagy, 1991; Bauer & Nation, 1993; Nerlich *et al.*, 2003; Sinclair, 2004; Stubbs, 2002; Wei & Light, 1973). In fact, this is the premise on which word families are based (Bauer & Nation, 1993). This may seem a logical approach for many native speakers and higher level ESL learners because, if asked to, they can define a high frequency word like *work*, based on what they deem as the basic or core meaning of a word. And they may even choose a

common meaning of the word work. Nevertheless, to ensure that an appropriate definition is given for specific circumstances, context must be known (*He worked his opponent in that soccer game; I must work hard to write a good paper; They work at Novell; The old woman slowly worked her way across the street*).

However, various researchers question the validity of a core or central meaning. Wei and Light (1973) point out that when words are grouped under head words, "there will be disagreement both on the criteria and on the result of using them" because the choice of those criteria are often arbitrary (p. 10). This can be seen just by comparing the entries of two or three different dictionaries or by comparing a word list based on word families versus lemmas. In addition, Anderson and Nagy's (1991) example of the word *give* above provides "specific evidence against the core meaning approach," because meaning and use are determined by distribution and context, thus making it difficult to determine which meaning is more core, basic, or central (p. 701). Which is the core idea of *give*, the idea of *donating* or the idea of *granting* or maybe the nuanced meaning of another related synonym? This is further evidence that there is a lack of precise correspondence between word form and meaning because both concepts and word forms sometimes overlap and interrelate or are completely separate and distinct, highlighting the idiosyncratic nature of vocabulary.

Knowing the meanings of words is a constant, life-long process for all L1 and L2 speakers of a language. Thus, learning and knowing the continuum of polysemous meanings as well as distinct homonymous meanings for the various word forms in English is a crucial part of language acquisition and development. Therefore, despite all of these aspects of language that make the defining of words complicated, lexicographers,

educators, and linguists continue to search for more psychologically and linguistically valid ways to define "words" or lexical items in English in order to facilitate language acquisition and language teaching. The development of tools such as the semantic-relatedness scale from Nagy and Anderson (1984) have been useful for those in the vocabulary field in terms of making more psychologically valid evaluations about word-forms and both their polysemous and homonymous meanings. (An example of this scale can be seen in Figure 1 of chapter 3). No doubt, subjectivity and disagreements in defining lexical boundaries will always exist due to the inherent nature of the form-meaning relationship and the infinite variety of human perceptions and experiences. Nevertheless, efforts to improve upon and solidify the construct of *word* and to develop more psychologically valid and thus pedagogically effective word lists and definitions need to continue.

In this pursuit to create an improved list, clear semantic boundaries between homographs and polysemes must be made, with the existing continuums of meaning, in order to define distinct homographs and facilitate the creation of a more psychologically valid high frequency word list. The present study attempts to define semantic boundaries for a small sample of homographs using Nagy and Anderson's (1984) scale of semantic relatedness. The primary purpose of doing this is to determine to what extent such homography exists and explore the possible implications that the findings could have on computer-generated high frequency word lists used for pedagogical purposes.

Problems with Form and Meaning: Defining the construct of word

Defining the construct of *word* is perhaps one of the biggest problems contributing to validity issues in electronically-based high frequency word counts and the

resulting word lists used for ESL teaching and learning purposes. Many linguists have found this to be a problem in creating and assessing vocabulary and word lists (i.e. Gardner, 2007; Nagy and Anderson, 1984; Read, 2000; Richards 1974; Wei and Light, 1973). Perhaps the most concise and comprehensive discussion to date on this specific problem is found in a recent paper by Gardner (2007). The primary purpose of his article is to "raise an awareness of this *Word* dilemma" and "to make recommendations for improving the validity of such research in informing English language education" (p. 242). Read (2000) also recognizes that this is the very crux of the vocabulary problem as far as making a "good quality list". He raises some important questions about the difficulties of defining what a word is for all practical intents and purposes. Read (2000) states that "it would require a substantial amount of skilled analysis and judgment to produce a good quality list, and so far no one has taken up the challenge" (p. 228).

Two of the three major validity issues Gardner (2007) raises in trying to define the construct of *word* are directly related to this study and essential issues in defining the gap of form-meaning relationships in computer-based word frequency lists. These two major issues (which he suggest are only superficially dealt with in applied corpus-based linguistics) include (1) the ability of ESL learners of varying skill levels and language backgrounds to make semantic connections between morphologically related words, and (2) the effects of polysemy and homonymy on L2 vocabulary acquisition (p. 243). The primary focus of this study is emphasis on the second issue. But the former must be addressed because of its direct connection with polysemy and homonymy.

The problem of existing polysemy and homonymy has already been dealt with extensively above. However, Gardner (2007) mentions one additional noteworthy

complication of this problem: computers lack the ability to disambiguate various forms (polysemous and homonymous) that cross between both word families and lemmas. This problem is illustrated in an analysis of the word forms *bear* and *bore*. He shows how sometimes they exist in the same semantic sphere, as polysemes (i.e. *bear/bore a burden*), while at other times they are definite homographs (i.e. *to bore a whole, the man is a bore, the bear slept in the cave, the child was born last night*), and concludes that "conceivably, a machine-based frequency count of word-family forms could link all of these forms of *bear* and *bore* together" (p. 251).

Gardner (2007) brings up another issue that is applicable to this study; clarification is needed about how language proficiency and background affect the way that words are psychologically connected both in form and meaning. This has a direct impact on psychological implications of homography as well as how the construct of *word* should be defined for word frequency counts. One of the primary concerns in creating word lists is whether words should be grouped by forms, lemmas, or word families, and whether multi-word lexical items should be counted as separate and distinct lexical items. Using word types (unique spellings) is generally ruled out because, as Gardner says, "it is highly unlikely that average readers in the third through the ninth grades . . . would see no connection between *boy* and *boys*" (p. 246).

Though neither the lemma nor the word family may be the ideal choice for defining the construct of *word*, these are the two primary ways in which words have been defined, grouped, and counted in word lists of the last 20 to 30 years. Thus, one of the key issues in determining how to count lexical items is ascertaining whether lemmas or word families are more psychologically valid. In light of the earlier points made about

semantic boundaries and the influence of L1 on ESL learners vocabulary acquisition, the decision is most certainly influenced by whether lists are made for native English speakers or for ESL learners.

Lemmas as a construct of word

In an analysis of the concept of lemma, Knowles and Mohd Don (2004) define a lemma as a unit based on a word form and all of its inflections, basically linking it to conventional parts of speech (POS). Often the concept of lemma includes a semantic value, making *bat* the animal and *bat* for baseball two separate lemmas. In the field of corpus linguistics the lemma has frequently been used "to generalize about the behaviour of groups of words in cases where their individual differences are irrelevant" (Knowles & Mohd Don, 2004, p. 70), thus making such word groups accepted as a manageable unit whose individual members share coinciding semantic and syntactic boundaries. Hence, for several computer-generated frequency counts, many linguists have chosen to group by lemma, probably under the assumption that inflectional forms of the same word will behave in identical ways and convey the same meaning. It also allows researchers to avoid tediously counting and analyzing each raw word form when some forms may almost completely overlap in meaning and usage.

However, in recent years some linguists have criticized the lemma as a justifiable unit of measurement for word frequency counts. One criticism is that when individual members of one lemma are analyzed in depth, they "can behave independently and develop their own meanings and collocations" (Knowles & Mohd Don, 2004, p. 71). For example, with the lemma DEAL (n), one would say '*it's no big deal*', but probably never '*it's no big deals*' or even '*they are no big deals*'. Perhaps a better example of this is the

lemma PROVIDE (v) in which the past participle form *provided* has developed a role as quite divergent from the rest of the members of the lemma, functioning at times as a subordinating conjunction, for example, *we can get the loan for the house provided that we have good credit* (Knowles & Mohd Don, 2004).

Another criticism made that is especially pertinent to this study is the fact that sometimes the blurred line between homonymy and polysemy makes the process of lemmatization messy, unsystematic, and subjective (Knowles & Mohd Don, 2004). As a result of this, researchers group forms into lemmas in different ways, according to varying criteria and lemmatization becomes less standardized and consequently less reliable for doing word frequency counts and other analyses. A good example of this is the following:

> the metaphorical use of *lion* (e.g. John is a lion) is likely to be treated as 'the same word', while the concrete and metaphorical uses of *crane* ('kind of bird' and 'machine for lifting heavy objects') are more likely to be treated as independent words and therefore members of different lemmas. (Knowles & Mohd Don, 2004, p. 70)

Some metaphorical senses have become so prominent that they are often distinguished as their own lemma by some lexicographers. However, the decision about where polysemy ends and where a distinct new meaning begins are not systematically created or easily agreed upon. This same point is made by Gardner (2007) about polysemous phrasal verbs (i.e. MAKE OUT - *makes out, making out, made out*; MAKE UP – *makes up, making up, make up*) (p. 244). He further points out that although lemmas theoretically attempt to distinguish between homographs, it is virtually impossible to account for them

on a practical level in the context of computer-generated frequency counts and word lists because no mega-corpora are semantically tagged and computers cannot yet adequately make accurate distinctions between homographs.

Word Families as a construct of word

Though the concept of a word family has been around for some time, Bauer and Nation (1993) re-established the concept by developing a systematic definition of word family that is easily standardized and operationalized. Perhaps due to this standardization of word families and with the development of the RANGE program (Heatley, Nation, & Coxhead, 2002) several researchers since then have used word families as the unit by which to define the construct of *word* in frequency counts. Word families are an excellent resource as a pedagogical tool for reinforcing and strengthening vocabulary knowledge and proficiency for both native speakers and ESL learners alike. Nation (2001b) points out that being aware of a core meaning and learning affixes can definitely make learning new forms of words in a word family easier. However, many concerns arise about the psychological validity of using the word family as a construct of *word*, particularly for word frequency counts and making word lists.

The primary issue concerning the psychological validity of word families is that they include so many forms under the guise of one meaning, consequently bringing out all of the problems listed with the lemma, but to an even more exaggerated level. The fundamental purpose for systematizing word family groups was "to set up a series of levels of affixes that could provide the basis for the staged systematic teaching and learning of these affixes for learners reading English" (Bauer & Nation, 1993, p. 254-255). Therefore, word families do not represent the words people know, or even the

psychological connections that the majority of people make between words. They essentially represent potential knowledge of words and their various related derivations and inflections. The varying word family levels could each be used to represent a distinct construct of *word* depending on learners' individual language proficiency.

Learners' knowledge of derivational relationships within word families is highly dependent upon the linguistic awareness and aptitude of the individual as well as what they have been exposed to, how they have been exposed to it, and how many times they have been exposed to it. The primary purpose of Bauer and Nation's study was not to form groups from which frequency counts could be made, but to "provide a consistent description of what should be considered to be part of a word family for readers at different levels of morphological awareness" (p. 255).

Additionally, Gardner (2007) identifies other concerns inherent in the nature of word families. One issue he brings up is the fact that the ways in which language learners (both L1 and L2) make associations between related word-forms in their mental lexicons remains unclear. A second is that the extremely productive nature of English affixes can produce a word family that has many derivational forms counted as one word. Other researchers who use word families admit that the development of affix knowledge is lengthy and continues through the teenage years and beyond (Bauer & Nation, 1993; see also Schmitt & Zimmerman, 2002).

The results of one study testing productive derivational knowledge of ESL learners suggest that advanced ESL learners do not have a productive knowledge of all derived forms of words for which they knew the base form (Schmitt & Zimmerman, 2002). And even the native English speakers in the control group had a "high but less

than complete productive knowledge of the derivational morphology" (p.160). This also raises questions about the claim by Bauer and Nation (1993) that "once the base word or even a derived word is known, the recognition of other members of the family requires little or no extra effort." It is true that their claims are based on receptive knowledge of readers and not productive knowledge. But care must be taken in the predictions and assumptions made about the types of semantic and morphological connections that learners will make. And a distinction must be made between the types of connections and abilities that native speakers will make compared to those that ESL learners will make.

Nation (2001b) also points out that "very frequent derived forms like *impossible* and *beautiful* are stored and used as if they were base words rather than being reconstructed according to derivational rules each time they are used" (p. 59). This is one more caveat to basing a system on the idea of a core or basic meaning.

Gardner (2007) makes one additional observation that is specifically related to homography and word frequency counts. He points out, as mentioned earlier with the example of *bear* and *bore*, that "potential meaning variation becomes even more convoluted when the morphological word family is considered" (p. 251). This is similar to the conclusion Ming-Tzu and Nation (2004) make in their study about homography in the Academic Word List (AWL). They suggest that "the use of word families inflates the problems caused by homography" because using a unit that includes less word types would separate out distinct meanings that "tend to be represented by different types" (p. 306). A further important difficulty to recognize is that, as with lemmas, in a computer-based frequency count, a modern computer cannot yet distinguish between homographs

to any acceptable level of accuracy. Anecdotal evidence of this is apparent on any translating program one can find on-line.

Finally, Gardner (2007) warns of three possible problems with computer-generated frequency counts and word lists that will persist if the form-meaning gap is not addressed: "(a) they will over estimate the true coverage of the word forms; (b) they will underestimate the actual user knowledge required to negotiate the word forms; and/or (c) they will underestimate the actual number of meanings inherent in the word forms" (p. 253).

As has been discussed here, all of the constructs of *word* that have been used for word counts (word forms, lemmas and word families) have weaknesses. None of them take into account multi-word lexical items as of yet, which is Gardner's (2007) third concern, and a vital one at that. Despite this weakness, of these three lexical grouping methods, the lemma is a more liberal approach than raw form frequency and a more conservative approach than combining all forms into a single word family. By using either single word-forms or word families for frequency counts, word lists lose some psychological validity for underestimating or overestimating the transparency of the form-meaning boundaries of related words, as Gardner (2007) suggests. Though the lemma is not ideal and operationalized in a systematic way like word families have been, it falls between these two extremes, making it the most balanced choice for an initial attempt to sample the lexical items in a large corpus of English.

Research Questions

This literature review highlights the significance of the research questions presented in this study:

1. What is the effect of a semantically-based versus a form-based analysis of word lemmas on the outcome of high-frequency English word lists generated by computers from electronic mega-corpora?
2. What are the implications of these lexical findings for:
    a. estimates of vocabulary coverage in texts (written and spoken)?
    b. the teaching and learning burden of vocabulary in ESL contexts?

CHAPTER THREE

Methods

Introduction

In order to determine the extent of semantic reliability and validity in computer-generated frequency word lists of English, this study analyzes randomly selected lemmas from the British National Corpus (BNC). Lemmas that occur at a pre-determined minimum frequency were selected and evaluated in order to 1) assess the extent of homography within each lemma form, and 2) determine what impact the number of homographs found within high frequency lemmas could have on the future of how word lists are compiled and used for pedagogical purposes. The following sections will describe the instruments that were used in this analysis as well as provide a detailed explanation of the processes and procedures that were carried out in order to explore possible answers, explanations, and/or clarifications for the questions posed in this study, which are the following:

1. What is the effect of a semantically-based versus a form-based analysis of word lemmas on the outcome of high-frequency English word lists generated by computers from electronic mega-corpora?
2. What are the implications of these lexical findings for:
    a. estimates of vocabulary coverage in texts (written and spoken)?
    b. the teaching and learning burden of vocabulary in ESL contexts?

Data Sources

There were two data sources used in this study: the British National Corpus and WordNet.

The British National Corpus (BNC)

The major data source used in this study is the British National Corpus (BNC). It is a collection of over 100 million words of written and transcribed spoken British English. It consists of 90% written materials and 10% recorded and transcribed spoken materials. It is generally accepted as a representative sample of a wide variety of written English input, and only a marginally representative sample of spoken British English (http://www.natcorp.ox.ac.uk/corpus/creating.xml).

One of the unique and valuable characteristics of this corpus is that it was grammatically tagged by CLAWS, meaning that each word has been marked with a grammatical value (e.g. part of speech/ grammatical function). The grammatical tagging was done by computer and then portions of the corpus were manually post-edited (http://www.natcorp.ox.ac.uk/corpus/creating.xml).

The BNC has some limitations. Perhaps the most obvious limitation is that it is solely British English. This affects lexical selection in important ways. First, there are words that are frequently used in British English (e.g. *mum*) that are seldom or never used in American or other dialects of English. Also, some word senses exist in British English, but are not used in American English and vice versa (e.g *pants* in British English is *underwear* and in American English is *trousers*). And finally, the BNC is now 15 years old and each year becomes more out-dated in its vocabulary selection. However, despite these limitations, the BNC remains an excellent tool and source for the study of the English language and one of the best of its kind in existence.

WordNet

WordNet was created by psychologist George A. Miller and his associates at Princeton University (http://wordnet.princeton.edu/). It is a well-respected and user-friendly on-line database that lists multiple senses of word forms accompanied by definitions, synonyms, and example sentences. The senses are categorized by part of speech (e.g. noun, verb, adjective, and adverb) and ordered according to sense frequency based on a semantic count from a small corpus (see Figure 1).

Figure 1 – Example screen shot of WordNet on-line

WordNet Search - 3.0 - WordNet home page - Glossary - Help

Word to search for: trump Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Noun

- S: (n) **trump**, trump card (a playing card in the suit that has been declared trumps) *"the ace of trumps is a sure winner"*
- S: (n) **trump** ((card games) the suit that has been declared to rank above all other suits for the duration of the hand) *"clubs were declared trumps"; "a trump can take a trick even when a card of a different suit is led"*
- S: (n) cornet, horn, trumpet, **trump** (a brass musical instrument with a brilliant tone; has a narrow tube and a flared bell and is played by means of valves)

Verb

- S: (v) **trump** (produce a sound as if from a trumpet)
- S: (v) outdo, outflank, **trump**, best, scoop (get the better of) *"the goal was to best the competition"*
- S: (v) **trump**, ruff (play a trump)
- S: (v) **trump**, trump out (proclaim or announce with or as if with a fanfare)

WordNet home page

As acknowledged by its authors, WordNet is not comprehensive in its list of senses. Since the present study focuses on counting occurrences of word meanings, supplementary senses were added to compensate for its omission of senses included in loosely or tightly bound idiomatic expressions, compound or multipart words, and phrasal verbs.

One other limitation or difficulty in using WordNet is that the example sentences that accompany each sense sometimes confused and hindered clarification instead of facilitating it. These confusing example sentences were especially problematic during the conflation process, which was performed in order to consolidate polysemous senses (unnecessary distinctions between two or more semantically related senses), (see Nagy and Anderson 1984). This may result from the small size of the corpus being used to create WordNet and to extract examples of word senses. Also, a mismatch between the British English-based corpus and the American English-based WordNet may have contributed to this limitation.

Instruments

VIEW Program

One instrument used in this study was the internet based program called VIEW (view.byu.edu) created by Mark Davies of Brigham Young University (now accessed at www.corpus.byu.edu/bnc). Essentially, Davies created an interface for the BNC, capable of allowing researchers to link specific lexical items to the natural spoken or written contexts in which the words originally occurred and thereby enable them to analyze the semantics of a word as well as discourse characteristics and patterns.

In this study, VIEW was primarily used to retrieve the target items *with* their surrounding contexts in a format that facilitated the semantic analysis of each instance of each word. It also provided the register information from which each context came as well as how many senses were listed for that particular word in WordNet. Examples of each of these will be shown in the General Procedures section.

Nagy and Anderson Semantic Relatedness Scale

Another instrument used in this study is a scale (see Figure 2) that was first published in a study done by Nagy and Anderson (1984) and has more recently been used in another related study (Ming-Tzu & Nation, 2004). This scale is used to determine levels of semantic relatedness of senses of a word in order to make defensible distinctions between those senses. This scale was chosen because it seems to adequately describe necessary levels of semantic relatedness. The descriptions are straightforward and clear,

Figure 2 - SCALE from Nagy and Anderson (1984)

Semantic Relatedness Level	**Description of the degree of relatedness**
0	**The meaning is the same as the base meaning.**
1	**The meaning is only slightly different from the base meaning.**
2 **(threshold for this study)**	**The meaning is related to the base meaning with some changes.**
3 **(threshold for Ming-tzu and Nation's study, 2006)**	**The meaning is substantially different from but is still related to the base meaning.**
4	**The meaning is very distantly related and almost totally different from the base meaning.**
5	**There is no relationship at all between this meaning and the base meaning.**

thus providing a manner in which to more reliably distinguish homographs of lemma forms.

The main limitation of this scale is that it requires some amount of subjective assessment. This subjectivity can be minimized by having multiple raters and discussions about rated items in which discrepancies are found. In the Nagy and Anderson (1984) study, ratings of 0–2 were determined as semantically transparent while ratings of 3–5 were classified as semantically opaque. Nagy and Anderson (1984) found that their raters were in agreement 76.6% of the time in determining transparency or opacity of the relatedness of two senses, which they suggested was "more than adequate" to support the arguments they made (p. 312). In Ming-Tzu and Nation's (2004) study, they changed the level of semantic transparency to ratings of 0–3 and classified meanings rated at a 4 or 5 as semantically opaque.

Procedures

Raters

Three raters determined semantic relatedness levels and aided in consolidating the senses of meanings that were identified as being semantically transparent. They also added some meanings that were deemed distinct from the existing WordNet senses. The three raters each have five or more years of teaching experience specifically in working with adult ESL students. Having experience teaching ESL students is relevant to this task because levels of semantic relatedness are most likely influenced by language background (Ijaz, 1986; Jiang, 2004; Zughoul, 1991), and English language learners do not always make the same types of connections between related word forms and meanings that native speakers do.

In addition to the three primary raters mentioned above, twelve additional raters completed the sense ratings, primarily working with double and triple ratings. Of the fifteen raters, two were linguistics professors at Brigham Young University, eight are current or former TESOL Masters students at Brigham Young University, one has her TESOL certificate, and an additional three raters have advanced degrees in unrelated fields. The one rater without an advanced degree has finished his BA in Russian studies and is starting medical school. Only four raters were not native English speakers but they are currently in the TESOL Masters program at BYU and have excellent English skills.

Word Selection

The first part of the process was the random selection of 100 high frequency words from the BNC for which a semantic analysis and semantic frequency count could be made. The lemma was chosen as the unit of measurement for the frequency list in this study. The first step was to create a lemmatized list of high frequency words from the BNC 100 million word corpus. From this list, all of the modals and function words of any type were eliminated in order to focus on words that possess semantic significance instead of those words which serve primarily grammatical purposes and functions.

Next, for statistical purposes, a minimum occurrence criterion of at least 1500 times in the written portion of the corpus and at least 500 times in the spoken portion of the corpus was set for high-frequency lemmas, from which 100 were randomly selected by a computer program to use for semantic analysis (see Appendix A for a complete list). Only the first 50 words were analyzed due to time constraints. Four of the 50 were found to be non-lexical items (e.g., *shall*), thus reducing the number to 46.

Because of general ambiguity of meanings when words are not in context (Sinclair, 2004), it was essential to have the words presented in their original contexts. A total of 200 randomly selected example contexts, 100 from written registers and 100 from spoken registers, were determined to be statistically sufficient (90% confidence) for counting the semantic frequency of a lemma in order to address the first research question as well as to facilitate a comparison between written and spoken semantic frequency, which is addressed in the second research question. It was determined that each context should include the 10–15 words of original context on either side of the target lemma to facilitate a more reliable rating of each word. These contexts were exported into an Excel file. The selected lemmas, starting with ACT (v) and ending with YESTERDAY (adv) were listed in alphabetical order, and for each selected word item the 100 spoken contexts were listed, followed by the 100 written contexts. Four example contexts (two spoken and two written) for the lemma FAIR (adj) taken from the excel file are shown here in Figure 3.

Figure 3 – Example contexts for the lemma FAIR (adj)

Word	Lemma	POS	KWIC (context)	Sense #	Register
Fair	Fair	AJO	assume we could afford it , he Just does n't like one in the bedroom so we have one downstairs . fair enough . Mhm . But other people erm , I , I think it 's a lot of money . it		S_brdcast _discussn
Fair	Fair	AJO	she ? No . I 'm earning it for her . I 'm earning it for her , that 's not fair . Let me ask Lindsey Here , Lindsey how 's your marriage going ? WHAT marriage ? Well tell me about		S_brdcast _discussn
Fair	Fair	AJO	" All we need is Quasimodo ." a moment later the door creaked open and he appeared , or a fair facsimile , a very Old man with grey hair down to his shoulders , a black dresscoat of velvet that had		W_fict_pro se
Fair	Fair	AJO	WHAT 's wrong with you , Celia ?" said Alan sharply . he was a tall , broad man with fair hair and clear hazel eyes . "anyone would think you do n't want Donna to have a bone-marrow transplant .		W_fict_pro se

Under the "register" category 'S' represents spoken and 'W' represents written followed by the micro register of the specific occurrence of the lemma in the BNC.

Sense Selection and Conflation

The next step in the process was to select an extensive and reliable list of senses for each word. The WordNet program was selected because of its reputation as an accessible and acceptable database of word meanings as explained above. Each of the 100 selected word items was looked up on the most current version of WordNet on-line and all of the senses for the corresponding lemma (usually coinciding with a part of speech) were copied onto a word document. For example, the senses of FAIR (adj) were listed on WordNet as follows:

Figure 4 – WordNet senses for the lemma FAIR (adj)

FAIR – adjective

1. S: (adj) **fair**, just (free from favoritism or self-interest or bias or deception; conforming with established standards or rules) *"a fair referee"; "fair deal"; "on a fair footing"; "a fair fight"; "by fair means or foul"*
2. S: (adj) **fair**, fairish, reasonable (not excessive or extreme) *"a fairish income"; "reasonable prices"*
3. S: (adj) bonny, bonnie, comely, **fair**, sightly (very pleasing to the eye) *"my bonny lass"; "there's a bonny bay beyond"; "a comely face"; "young fair maidens"*
4. S: (adj) **fair** ((of a baseball) hit between the foul lines) *"he hit a fair ball over the third base bag"*
5. S: (adj) average, **fair**, mediocre, middling (lacking exceptional quality or ability) *"a novel of average merit"; "only a fair performance of the sonata"; "in fair health"; "the caliber of the students has gone from mediocre to above average"; "the performance was middling at best"*
6. S: (adj) **fair** (attractively feminine) *"the fair sex"*
7. S: (adj) clean, **fair** ((of a manuscript) having few alterations or corrections) *"fair copy"; "a clean manuscript"*
8. S: (adj) honest, **fair** (gained or earned without cheating or stealing) *"an honest wage"; "an fair penny"*
9. S: (adj) **fair** (free of clouds or rain) *"today will be fair and warm"*
10. S: (adj) **fair**, fairish ((used of hair or skin) pale or light-colored) *"a fair complexion";*

(from the WordNet program available on-line at wordnet.princeton.edu)

Once the list of each word and all of its senses was compiled, it was necessary to conflate some of the meanings that were determined to be semantically similar in order to only have senses that would be considered separate lexemes and homographs within each lemma form. Nagy and Anderson's (1984) scale (see figure 2) was used primarily as a guideline for the three raters as they negotiated through the conflation process. The raters conflated separately at first, and later came together to come to a consensus about the

Figure 5 – Conflated senses of FAIR (adj)

FAIR – adjective

1. S: (adj) **fair**, just (free from favoritism or self-interest or bias or deception; conforming with established standards or rules) *"a fair referee"; "fair deal"; "on a fair footing"; "a fair fight"; "by fair means or foul"* + S: (adj) honest, **fair** (gained or earned without cheating or stealing) *"an honest wage"; "an fair penny"*
   NOTES: related to definition #4 in the sense of conformity to rule; 'fair enough' = just, alright, acceptable, good, fine; fair play (within the rules)
2. S: (adj) **fair**, fairish, reasonable (not excessive or extreme) *"a fairish income"; "reasonable prices"*;
   NOTES: can mean more towards a lot or toward a large amount but not to the complete extreme or excessiveness
3. S: (adj) bonny, bonnie, comely, **fair**, sightly (very pleasing to the eye) *"my bonny lass"; "there's a bonny bay beyond"; "a comely face"; "young fair maidens"* + S: (adj) **fair** (attractively feminine) *"the fair sex"*
   NOTES: related to definition 10 (CS8) depending on cultural norms
4. S: (adj) **fair** ((of a baseball) hit between the foul lines) *"he hit a fair ball over the third base bag"*
   NOTES: somewhat related to 1 in conformity of rules
5. S: (adj) average, **fair**, mediocre, middling (lacking exceptional quality or ability) *"a novel of average merit"; "only a fair performance of the sonata"; "in fair health"; "the caliber of the students has gone from mediocre to above average"; "the performance was middling at best"*
   NOTES: somewhat related to definition 2 in being in the middle rather than at the extremes
6. S: (adj) clean, **fair** ((of a manuscript) having few alterations or corrections) *"fair copy"; "a clean manuscript"*
7. S: (adj) **fair** (free of clouds or rain) *"today will be fair and warm"*
8. S: (adj) fair, fairish ((used of hair or skin) pale or light-colored) *"a fair complexion";*

discrepancies among their ratings. This process helped to verify true lexemes that could then be used for the semantic tagging and frequency counts. The reduced list created during this process was used for the final semantic ratings. Occasionally, senses were added to the list after ratings began when additional meanings (not in WordNet) were found in the contexts. Re-evaluations were made of previous ratings where appropriate.

Figure 5 above shows the final list of senses for FAIR (adj), including notes from the conflation process. It is important to notice in the above example that two definitions were eliminated by the conflation process (i.e. FAIR (adj) began with ten definitions that were reduced to a list of eight definitions through the process described above).

In this example of FAIR (adj), it is evident that some of the definitions are related yet remain distinct. For example, one could argue that the meaning 'very pleasing to the eye; attractively feminine' is related to the meaning 'fair skinned' because light-colored skin was once considered very attractive and feminine. Or one could argue a somewhat transparent semantic relationship between the meaning 'not excessive or extreme' and the sense for weather 'free of rain or clouds'. The decision to conflate such senses or to keep them separate was based on the definitions of the Nagy and Anderson scale as well as consideration of the following: 1) how psychologically closely they were connected, 2) how challenging it was to make a connection between the figurative usages and the literal ones, and 3) how frequently both the figurative and literal senses were used (based on WordNet frequency counts).

This process produced varied results. Some words had many senses combined into just a few while others stayed relatively the same with little or no conflations. And in some cases, more senses were added to the original list when contexts were

encountered in which it was apparent that a lemma represented a distinct sense that was not listed among the WordNet senses.

The sense conflation process was very challenging and required that each word be assessed by at least two of the three raters. Disagreements on semantic relatedness were resolved by discussion between the raters on a word-by-word basis.

Assigning Senses to Lemmas in Context

After the senses of the words were conflated, the lemmas with their 200 contexts were each randomly ordered. Then the raters used a final, conflated sense list to assign an appropriate sense to each individual example context (see Figure 6). A double rating was done on all 46 lemmas and their 200 contexts, and when necessary a triple rating was performed to assure reasonable accuracy of sense assignment. Due to time constraints, only 46 lemmas of the original 100 were analyzed. Originally, the first 50 lemmas on the

Figure 6 – Semantic ratings of FAIR (adj)

Word	Lemma	POS	KWIC (context)	**Rating (Sense #)**	Register
Fair	Fair	AJO	assume we could afford it , he Just does n't like one in the bedroom so we have one downstairs . fair enough . Mhm . But other people erm , I , I think it 's a lot of money . it	**1? Expression 'fair enough'**	S_brdcast_ discussn
Fair	Fair	AJO	she ? No . I 'm earning it for her . I 'm earning it for her , that 's not fair . Let me ask Lindsey Here , Lindsey how 's your marriage going ? WHAT marriage ? Well tell me about	**1**	S_brdcast_ discussn
Fair	Fair	AJO	“All we need is Quasimodo .” a moment later the door creaked open and he appeared , or a fair facsimile , a very Old man with grey hair down to his shoulders , a black dresscoat of velvet that had	**2**	W_fict_pro se
Fair	Fair	AJO	WHAT 's wrong with you , Celia ? “; said Alan sharply . he was a tall , broad man with fair hair and clear hazel eyes . “anyone would think you do n't want Donna to have a bone-marrow transplant .	**8**	W_fict_pro se

list were selected from the list. However, four additional lemmas (HERE, WHERE, AGAIN, and SHALL) were eliminated because they were non-lexical items, either determined as mis-tagged in the BNC, or not filtered correctly by the computer program.

Because of computer limitations some contexts were repeated twice. If this occurred, an X was placed in the sense # box, and these contexts were simply deleted in the final sense counts. Lemmas that had been mis-tagged in the BNC were marked with a 99. In a few instances a lemma occurred as a proper noun that was not listed in the final sense list and was marked with a PN for *proper noun*.

Another important decision dealt with the fact that some nouns were actually often a part of a compound noun. For example, the lemma *school* usually occurred with another word to complete an idea: *school* age, *school* report, *school* building, *school* system, *school* teachers, and *school* districts. So instead of counting the word *school* as an adjective and marking those instances with a 99 (i.e. a mis-tagged item), they were counted as nouns and marked according to the appropriate sense of the word, with a notation made for a *compound noun* (CN). In addition, lemmas appearing in an *idiomatic expression* (IE), *phrasal verb* (PV), or *compound verb* (CV) were appropriately noted.

First and second ratings were compared, and triple ratings were performed, when necessary, with the third rater simply choosing between the different ratings of the first two raters. Once all of the triple ratings were finished, the senses were counted with Excel to determine how frequently each sense occurred in total (written plus spoken) as well as separately in the two registers. Because some contexts could not be used, comparative analyses were performed on percentages rather than raw frequencies.

CHAPTER FOUR

Results and Discussion

Introduction

The purpose of this study is to examine more closely the theoretical and methodological practices of computer-based word frequency counts from which word lists are created. Specifically, the study attempts to investigate a primary concern that has been made by various linguists about the problem of evaluating word form instead of word meaning (Coniam, 1999; Gardner, 2007; Read, 2000), particularly among the highest frequency word forms of the language. Since word lists first became prevalent in the mid 1950's, compilers have mostly ignored the existence of homographs in the high-frequency word forms from computer-based word lists. This chapter presents the results and some direct implications of a semantically-based frequency count performed on a representative sample of lemmas in context from the BNC. The results and discussion are guided by the following research questions:

1. What is the effect of a semantically-based versus a form-based analysis of word lemmas on the outcome of high-frequency English word lists generated by computers from electronic mega-corpora?
2. What are the implications of these lexical findings for:
    a. estimates of vocabulary coverage in texts (written and spoken)?
    b. the teaching and learning burden of vocabulary in ESL contexts?

In brief review, 100 lemmas were randomly selected from a lemmatized frequency count of the BNC done by Mark Davies at BYU. This number was reduced to the first 46 on the list for the final semantic frequency analysis. For each of these 46

lemmas, 200 examples (100 spoken and 100 written) were analyzed, and the senses were counted and totaled for each lemma. Some of the words ended up with less than 200 examples because of mis-tagging or repeated contexts. This was taken into account by using percentages for comparative analysis purposes.

Results of a lemmatized frequency count vs. a lexeme-based frequency count

In this study, a criterion was set that any sense accounting for 10% or more of the rated occurrences (representativeness) of the lemmas would be identified and counted as a separate lexeme on a high frequency word list. Thus, any lemma with two or more senses accounting for at least 10% of the occurrences was considered to have homography. In other words, the lemma forms in current word lists represent various different lexemes of the language. The existence of homography within lemma forms implies that both the content of word lists and the rankings of items on that list would be likely to change in a semantically-based frequency word list, thus suggesting that the current lists may not give an accurate view of English word frequencies.

The following example illustrates this point. In a lemmatized frequency list from the BNC, created by Adam Kilgariff (found at http://www.kilgarriff.co.uk/bnc-readme.html), the lemma FAIR (adj) easily makes the top 2000 list of high frequency lemmas, at number 1393 and it occurs 6936 times in 100,000,000 words. However, in a lexeme-based list, a possibility of eight separate lexemes (according to the conflated senses list of the lemma form FAIR (adj) shown in chapter 3) could exist within those 6936 occurrences of the lemma form. This would potentially divide that number into lower total frequencies because each homograph would represent a separate lexeme, which includes both form and meaning. As a result, the lexeme [FAIR] (adj), meaning *not*

excessive or extreme and the lexeme [FAIR] (adj), meaning *unbiased or just*, would both be represented separately on a lexeme-based list and, therefore, divide the 6936 occurrences into several smaller numbers representing the frequencies of individual lexemes. Thus, their total frequencies would be smaller, and their overall frequency rankings would be lower.

Interestingly, in this study, exactly half of the 46 lemmas analyzed revealed having two or more senses that fell within the set criterion of at least 10% of the total occurrences. A compilation of all of the data for each of the 46 lemmas can be found in Appendix A. One of the tables in Appendix A shows the amount of homography that exists within each lemma form, and the proportional representations of each of the senses for these lemmas. One interesting observation from this study was that some of the senses from the conflated lists were not represented at all in the 200 randomly selected sentences from the BNC. This could either mean that those senses are not very frequent, or that they are only frequent in very specific registers. Appendix A shows the total number of possible senses as well as the number of senses that were actually represented in the random sample from the BNC.

In analyzing the results, the five lemmas with the highest amounts of homography are WORK (v), CHARACTER (n), BUSINESS (n), MEMORY (n), and ACT (v). Table 1 shows a breakdown of the proportional representations of each sense for each of these five lemmas. It is very apparent in looking at Table 1 that the percentages for each of these lemmas show a noticeable spread of representation across their respective senses.

To explore the implications of these findings with regard to the first question in this study, the results for the lemma WORK (v) will be examined as the first example. On the final list of conflated senses, WORK (v) ended up with 16 total senses. But

Table 1 – The 5 lemmas exhibiting the greatest amount of homography

	WORK (v)	CHARACTER (n)	BUSINESS (n)	MEMORY (n)	ACT (v)
Sense 1	**27%**	**32%**	**32%**	**36%**	**41%**
Sense 2	**31%**	**32%**	**35%**	**30%**	**41%**
Sense 3	**20%**	**13%**	**8%**	**35%**	**18%**
Sense 4	**2%**	**14%**	**24%**	**-**	**0%**
Sense 5	**3%**	**9%**	**1%**	**-**	**-**
Sense 6	**1%**	**2%**	**0%**	**-**	**-**
Sense 7	**0%**	**0%**	**1%**	**-**	**-**
Sense 8	**0%**	**-**	**-**	**-**	**-**
Sense 9	**1%**	**-**	**-**	**-**	**-**
Sense 10	**1%**	**-**	**-**	**-**	**-**
Sense 11	**14%**	**-**	**-**	**-**	**-**
Sense 12	**0%**	**-**	**-**	**-**	**-**
Sense 13	**1%**	**-**	**-**	**-**	**-**
Sense 14	**0%**	**-**	**-**	**-**	**-**
Sense 15	**1%**	**-**	**-**	**-**	**-**
Sense 16	**1%**	**-**	**-**	**-**	**-**

only 12 different senses are represented in the samples from the BNC. Four of these senses (senses 1, 2, 3, and 11) are represented at the criterion level of 10% or more of the total rated occurrences of the lemma WORK (v). Table 2 shows the projected impact of this semantically-based frequency count on a high frequency word list by comparing a lemmatized count with a lexeme-based count. It shows that by taking into account all of

the homographs with a substantial amount of representativeness in these 23 lemmas, the number of vocabulary items on the word lists would increase from 23 lemmas to 58 lexemes – more than doubling the number of separate words that should be included in a frequency count.

Table 2 – Contrast of a lemma count before and after distinguishing homographs

Lemma	# of Lemmas (Form Analysis)	# of Lexemes (Semantic Analysis)
WORK (v)	1	4
CHARACTER (n)	1	4
BUSINESS (n)	1	3
MEMORY (n)	1	3
ACT (v)	1	3
BATH (n)	1	3
DIRECTION (n)	1	2
APPLICATION (n)	1	3
STAND (v)	1	2
DEVELOP (v)	1	2
WELL (adj)	1	2
ANSWER (n)	1	2
GAS (n)	1	3
SUBJECT(n)	1	3
FAIR (adj)	1	2
SHOW (v)	1	2
AWFUL (adj)	1	2
GREEN (adj)	1	3
PULL (v)	1	2
BUSY (adj)	1	2
BACK (adv)	1	2
MATCH (n)	1	2
REST (n)	1	2
TOTAL	**23**	**58**

In order to more fully assess the effect that this kind of homography would have on the rank of each lemma in computer-generated word list, it is important to look at a comparison of rankings of a lemmatized frequency list and of a lexeme-based frequency list. Rankings have been extrapolated based on the results of the semantic frequency

analysis done in this study. Table 3 shows a comparison of word distributions for these two types of approaches to counting frequency, using the lemmas WORK (v) and CHARACTER (n). The lemma WORK (v), in Kilgariff's lemmatized frequency count, occurs 67,842 times in the BNC and ranks as the 129th lemma in a frequency list that includes all of the

Table 3 – Frequency results based on Form Frequency vs. Semantic Frequency

	Lemmas		**Lexemes**		
Word	**Lemma Freq***	**Lemma Rank**	**Sense #**	**Sense Freq****	**Sense Rank**
WORK (v)	67,842	129	2	21,133	488
			1	18,168	573
			3	13,345	761
			11	9,640	1,046
			Other	5,556	***
			Total	67,842	
CHARACTER (n)	12,511	818	2	4,004	2,209
			1	3,941	2,236
			4	1,689	3,937
			3	1,564	4,146
			Other	1,313	***
			Total	12,511	

* Frequency based on Kilgariff (website created 20 Nov. 1995)
**Projected frequencies based on manual sense sampling
***Undetermined

common function words such as articles, modals, pronouns, and prepositions. But if each sense were separated out as a distinct homograph, and counted as individual lexemes, each of the four lexemes [WORK] (v) would drop considerably in rank on the word list. Sense 2 of WORK (v) is the largest proportion of the lemma (31.15%) and could be projected to occur around 21,133 times. This would move the *most* frequent sense of

WORK (v) to the rank of 488, more than two times lower than its original rank when sense was not considered.

The projections of the remaining senses of WORK (v) are presented in Table 3. Sense 1, which represents 26.78% of the occurrences of the lemma, would occur 18,168 times and rank at 573; sense 3 would occur 13,345 times and rank at 761; and sense 11 would occur 9640 times and ranks at 1046. Of course the remaining 5556 occurrences would spread across the remaining 8 senses represented in the count. From this division, it seems that a semantically-based frequency count could potentially have a tremendous effect on a high-frequency word list by 1) adding many new lexemes, 2) decreasing the actual number of occurrences of many lemmas, and 3) shifting all lexemes to different ranks on the list.

One important point to make here is that these projected rankings in Table 3 would only hold true if each lemma in this study were the only lemma in the list which had a sizeable amount of homography within the lemma form and therefore would be the only lemma that would affect rankings. With more lexemes added in, the rankings would likely be even more substantially affected. Even though 50% of the lemmas in this study did not show any significant amount of homography, there was still an increase of 35 new lexemes to add to a word list, almost doubling the 46. If the inclusion of meaning to characterize word units has this large of an effect on such a small random sample, the possible effects on a larger scale could be quite substantial. Even by assessing the amount of homography in the top 2,000 most frequent words, if a similar amount of homography were found, the content and order of a frequency list would vary

considerably. And this is without considering multi-part words, such as idioms and phrasal verbs, as separate lexemes.

There are two additional important points to make about the projections for the lemma WORK (v) from Table 3. First, it is important to note that because this lemma is an extremely high frequency lemma, all of its homographs stay well within the top 2,000, and three of the four stay within the top 1,000 most frequent lemmas. Thus, four of the senses could possibly be considered highly frequent English lexemes. Again, these rankings would most likely be considerably altered when all homographs from high frequency lemmas are taken into account in such a list. In contrast to the lemma WORK (v), the second lemma, CHARACTER (n), moves from ranking in the top 1,000 most frequent words, at 818, to all four of its homographs dropping below the top 2,000. However, as was mentioned in chapter 1, the definition of *high frequency words* may need to be expanded to include more than 2,000 lexemes.

Table 4 compares the top ten items from the original 46 lemmas on lemmatized and lexeme-based lists side by side. (See the complete lists in Appendix B.) The left side of the table shows the frequency counts and rankings from Kilgariff's lemmatized list. The right side shows how the list changes when lexemes are counted. The table demonstrates how an actual word list would be altered when semantically-based counts occur and highlights an important point. The rankings decreased for all of the lexemes, even those that had little or no homography, simply because identifying multiple homographs increased the total number of words on a frequency list.

Table 4 – Results of lemmatized and lexeme-based frequency counts

	Lemmas	Lemmatized ranking	Lemmatized frequency count	Lexemes	Lexeme ranking	Lexeme frequency count
1	BACK (adv)	118	75,494	[BACK- 1] (adv)	162	58,719
2	WORK (v)	129	67,842	[SCHOOL] (n)	186	50,660
3	SHOW (v)	163	58,152	[TRY] (v)	189	50,068
4	TRY (v)	174	54,422	[SHOW- 1] (v)	234	39,886
5	SCHOOL (n)	181	52,227	[POLICE] (n)	360	27,508
6	BUSINESS (n)	244	38,204	[NAME] (n)	346	28,432
7	STAND (v)	285	32,899	[DIFFICULT] (adj)	466	22,033
8	NAME (n)	292	32,309	[WORK- 2] (v)	488	21,133
9	SUBJECT (n)	336	29,091	[ROAD] (n)	488	21,024
10	POLICE (n)	360	27,508	YESTERDAY (adv)	542	19,070

(based on analysis of all 46 lemmas; from most to least frequent)

In order to understand the full implications of these findings, it is also important to explore the results of the lemmas at the other end of the spectrum (i.e., the list of lemmas displayed in Table 5 which show little or no homography). The frequency of these 23 lemmas essentially remains unchanged. The first 12 revealed no homography at all, suggesting that there is no lemma-lexeme distinction. The remaining 11 exhibited minimal homography; however, only one sense for each met the 10% criterion for representation. As shown in Table 5, the lemma frequency and rank of the first 12 remain exactly the same as those for a lexeme-based count, while the frequencies and ranks of the final 11 change minimally when they are divided into lexemes.

Table 5 – Frequencies and rankings of lemmas showing little or no homography

Word	Lemma Freq*	Lemma Rank	Lexeme Freq**	Lexeme Rank
POLICE (n)	27,508	360	Same	Same
DIFFICULT(adj)	22,033	466	Same	Same
NECESSARY (adj)	18,107	573	Same	Same
RESPONSIBILITY (n)	12,078	846	Same	Same
TRAIN (v)	11,907	855	Same	Same
UNIVERSITY (n)	11,367	893	Same	Same
CLOTHES (n)	7,308	1,331	Same	Same
MIDDLE (n)	6,363	1,509	Same	Same
WHILE (n)	6,058	1,578	Same	Same
PROPERLY (adv)	5,667	1,680	Same	Same
CONGRESS (n)	5,544	1,708	Same	Same
PRESUMABLY (adv)	3,279	2,532	Same	Same
TRY (v)***	54,422	174	50,068	190
SCHOOL (n)***	52,227	182	50,660	187
NAME (n)***	32,309	292	28,432	347
ROAD (n)***	23,103	441	21,024	489
YESTERDAY (adv)***	19,459	533	19,070	543
PAGE (n) ***	14,546	708	14,110	735
START (n) ***	9,268	1,083	9,083	1,111
NORTH (n) ***	8,949	1,123	8,681	1,162
MARRY (v) ***	8,631	1,171	8,545	1,187
SLEEP (v)***	6,992	1,381	6,782	1,427
MUM (n)	+	+	+	+

* Frequency based on Kilgarriff (website created 20 Nov. 1995)
**Projected frequencies based on manual sense sampling from this study
***Other senses were represented very infrequently, thus, altering the Kilgarriff numbers
+did not show up on Kilgarriff's list

Table 6 illustrates the idiosyncratic nature of vocabulary. All four parts of speech are represented in both lemmas showing substantial homography and the lemmas with little or no homography. Thus, part of speech cannot be used to consistently predict the likelihood of homography in a lemma. In looking at this table, perhaps the one part of speech that shows the highest possible degree of predictability is the adverb, which seems to exhibit notably less representation in the group of homonymous lemmas (only 25%). However, the sampling in this study is too small to draw any decisive

Table 6 – Homography in the different parts of speech

POS	Total # of 46 randomly selected lemmas	Lemmas with **NO** criterion homography	Lemmas with substantial homography
Adjectives	7	2	5
Adverbs	4	3	1
Nouns	25	13	11
Verbs	10	4	6

conclusions. Also, both lists have lemmas that vary in rank from the top 100s up into the 2500s. (WELL (adj) is the one exception of the 46 lemmas showing homography that falls in a lower frequency band, ranking at 3247). This seems to suggest that there is no clear defining characteristic or level of predictability for what parts of speech will tend to have homography or not. It most likely requires analysis of lemmas more extensively and on an individual basis.

The one factor that would seem an obvious predictor of homography is the number of possible senses that WordNet lists for each lemma. Logically, the lemmas with more listed meanings from WordNet have a greater probability of exhibiting some homography. After consolidating polysemous senses and adding additional homographs to the WordNet senses, the average number of senses from WordNet for the group that

showed notable amounts of homography was just over 6 senses. The lemmas with no homography averaged 2.4 senses.

However, there are exceptions to this rule on both ends of the spectrum, once again demonstrating the idiosyncratic nature of vocabulary. For example, the lemma TRAIN (v) has five total distinct senses, yet only one was represented in the 200 examples of TRAIN (v) in the BNC. By contrast, several of the lemmas that ended up with no apparent homography had several senses represented in the BNC examples. The most extreme example is the lemma TRY (v) which has 8 possible senses and only one sense meeting the coverage criterion of 10% of the total occurrences. Other examples are NAME (n) with 5 senses and NORTH (n) and PAGE (n) with 4 senses each, with none of the newly distinguished lexemes reaching the 10% representation criterion in the 200 samples. It is quite possible that these senses would be represented to a greater extent in a larger sample size, though the probability of those meanings becoming prominent enough to reach the 10% representation criterion is low. (See Appendix A as a reference.)

In contrast, several lemmas that have only two or three possible senses had both lexemes highly represented in the BNC samples. For example, ANSWER (n) has only two senses, both of which are highly represented in the BNC ([ANSWER-1] (n) = 53% and [ANSWER-2] (n) = 47%) and consequently need to be represented in a word list as separate lexemes. In addition, 5 other lemmas each have 3 possible senses with at least two of them representing at least 30% or more of the 200 samples. On the other hand, 3 lemmas from the list that had no notable homography had 3 possible senses of which none had a minimum of 10% coverage. Thus, knowing how many possible senses there are is not always helpful in predicting which ones *will*, in reality, have substantial homography in a

representative sample of English, illustrating the idiosyncratic nature of words. (See Appendix A as a reference.)

Effects of lexeme-based frequency counts on estimates of vocabulary coverage in texts

Word coverage in texts has been a topic of great interest in the field of applied linguistics (specifically TESOL) in recent years. For example, the GSL is considered to cover around 80% of running words in texts in English. Nation (2006) defines text coverage as "the percentage of running words in the text known by the readers" (p. 61). Because the knowledge of human subjects is not used in this word analysis, it is not possible to use this definition in reporting the results with regard to coverage. Thus, coverage, in a general sense, will be defined as “the percentage of the running words in a text or corpus that are also in, or covered by, a particular word list” (Nation & Kyongho, 1995, p. 35). Primarily, the goal is to define how much of a text a certain group of words (i.e. a word frequency list) covers.

Even before looking at homography, estimates of word coverage are largely influenced by how a word is measured, as was discussed in depth in chapter 2. This study specifically looks at lemmas. However, in looking at coverage it is important to compare the three major constructs of *word*: word families, lemmas, and lexemes. Approximate frequencies of word forms (representing the idea of word families here) were calculated by adding all of the possible inflections and derivations included in the various parts of speech (according to Kilgariff’s list). These total form frequencies are compared with the lemma frequencies in Table 7. In these approximations, if the parts of speech are not distinguished, the word form (representing the concept of word family) *work*, with all of its inflections and derivations, occurs 130090 times. If the occurrences

of only the lemma WORK (v) are counted, this divides the number of occurrences almost in half at 67842 times. Table 7 shows frequency counts based on which construct of *word* is used: lemmas or word forms. These numbers suggest that the higher the form frequency count, the greater the coverage will be in a text because it covers all forms of a word. This implies that if L2 learners know the word form *work* and its core meaning, they will also know and recognize all 130090 occurrences of the various forms of *work* and all of their meanings and uses. With regard to the lemmas WORK (v) and WORK (n), L2 learners would have to be familiar with both the noun-forms and the verb-forms in order for the lemmas to cover the same amount of text as the word-form construct.

Table 7 - Contrast between a lemmatized and a form frequency count

Lemma	Lemmatized Freq. Count	Form Freq. Count	POS Included in Form Freq. Count
NAME (n)	32309	38593	noun, verb
WORK (v)	67842	130090	noun, verb
SCHOOL (n)	52227	52227	noun
WELL (adj)	2241	145852	adjective, adverb, interjection, noun
GREEN (adj)	9013	12183	adjective, noun
DIFFICULT (adj)	22033	22033	adjective
SUBJECT (n)	29091	32392	noun, adjective, verb
TRY (v)	54422	55799	verb, noun
MARRY (v)	8631	8631	verb
UNIVERSITY (n)	11367	11367	noun
YESTERDAY (adv)	19459	19459	adverb
BUSY (adj)	5221	5221	adjective
WHILE (n)	6058	56606	noun, conjunction
MUM (n)	Not listed	Not listed	noun, adjective
ROAD (n)	23103	23103	noun
RESPONSIBILITY (n)	12078	12078	noun
START (n)	9268	50297	noun, verb
CONGRESS (n)	5544	5544	noun
MATCH (n)	8718	14626	noun, verb
BACK (adv)	75494	106315	adverb, adjective, noun, verb
GAS (n)	8133	8133	noun
FAIR (adj)	6936	9635	adjective, adverb, noun
PRESUMABLY (adv)	3279	3279	adverb
REST (n)	14440	19146	noun, verb
SLEEP (v)	6992	10624	verb, noun
PAGE (n)	14546	14546	noun
AWFUL (adj)	2960	2960	adjective

-See Appendix D for this complete table including all of the 46 lemmas
-See Table 3 or Appendix B for examples of a lexeme-based frequency counts

As was mentioned above, in this study 23 of the 46 words analyzed showed extensive homography according to the 10% representation criterion set. In looking at some specific examples, this point can be made more clearly. Continuing with the example lemma WORK (v), it has a total count of 67842 in the BNC. If it is divided proportionally (as shown in Table 3) according to the homographs that were found, each lexeme would cover a smaller percentage of a total text or corpus. ESL students would need to know the four lexemes of [WORK] (v) that represent a large enough proportion of the lemma (10% or more) in order for their coverage of that word-form to be at 98%. These are the following senses of WORK (v): sense 1 = exerting or causing others to exert energy by doing mental or physical work for a purpose or out of necessity", sense 2 = to be employed/ have an occupation, sense 3 = to cause to operate, function or have an effect or outcome, and sense 11 = to work something out or solve it or work through it (from WordNet definitions). Thus, if a reader only knew 2 of those senses and three out of the four all remained in the top 1000 most frequently occurring lexemes, this would mean that a learner would have lower comprehension than expected due to lower coverage knowledge.

Comparison of written vs. spoken coverage

Another important aspect of the second research question in this study that needs to be addressed here is the impact of a semantically-based count on coverage in written versus spoken registers. Research on individual lexical items has consistently shown that register and range have a considerable impact on the results of vocabulary frequency counts. The results of this study support the results of this research and observations made in this area as well. Essentially, in the 23 lemmas with substantial homography,

there were 31 instances found in which a specific sense had considerable disparity in the frequency of use in either the written or the spoken register. In 10 of these the disparity is 20% or more. Table 8 presents these lemmas and the senses showing notable disparities between the two registers. Again, the threshold of significance was set at 10%. So any sense that showed at least a 10% disparity in usage between the two major

Table 8 – Lexemes with substantial sense disparities in spoken vs. written registers

Lexeme	% Spoken coverage	% Written coverage	% differential
[WORK- 11] (v)	20	8	12
[CHARACTER- 1] (n)	23	40	17
[CHARACTER- 2] (n)	39	25	14
[BUSINESS- 2] (n)	26	43	17
[BUSINESS- 4] (n)	30	17	13
[MEMORY- 3] (n)	41	28	13
[ACT- 2] (v)	49	35	14
[ACT- 3] (v)	11	23	12
[BATH- 2] (n)	56	39	17
[APPLICATION- 2] (n)	61	37	**24**
[APPLICATION- 4] (n)	15	33	18
[DEVELOP- 1] (v)	42	60	18
[DEVELOP- 2] (v)	55	40	15
[WELL- 2] (adj)	40	59	19
[ANSWER- 1] (n)	38	63	**25**
[ANSWER- 2] (n)	69	31	**38**
[GAS- 1] (n)	16	40	**24**
[GAS- 5] (n)	72	47	**25**
[SUBJECT- 1] (n)	75	53	**22**
[SUBJECT- 2] (n)	16	26	10
[SUBJECT- 4] (n)	4	19	15
[AWFUL- 1] (adj)	60	80	**20**
[AWFUL- 2] (adj)	40	20	**20**
[GREEN- 1] (adj)	66	78	**22**
[BUSY- 1] (adj)	84	66	18
[BUSY- 2] (adj)	15	34	19
[BACK- 1] (adv)	88	68	**20**
[BACK- 2] (adv)	7	23	16
[MATCH- 1] (n)	16	6	10
[REST- 1] (n)	94	81	13
[REST- 2] (n)	6	17	11

registers was included in these 31 instances.

The three lexemes showing the most extreme contrast between the two registers will be looked at in more detail. First, the lexeme [ANSWER- 1] (n), meaning *a response or reply as in a speech act* and the lexeme [ANSWER- 2] (n), meaning *a solution or result* exhibit fundamentally different patterns of frequency in written and spoken registers. The semantic frequency count showed that [ANSWER- 1] (n) is used 38% more often in spoken communication than in written while [ANSWER- 2] (n) is used 25% more often in written than in spoken. The comparison of the lexemes [GAS- 1] (n), *a gaseous state (as opposed to liquid or solid)* and [GAS- 5] (n), *natural gas or specifically fossil fuel in a gaseous state* is another good example of the disparity between registers. [GAS- 1] (n) is used 24% more often in written while [GAS- 5] (n) is used 25% more in spoken. The final examples are the lexemes [AWFUL- 1] (adj), *something bad, displeasing, mean or offensive that causes fear or dread, etc.*, and [AWFUL- 2] (adj), *extreme in degree, extent, impact or amount*. [AWFUL- 1] (adj) is used 20% more frequently in written while [AWFUL- 2] (adj) is used 20% more frequently in spoken registers. These findings indicate that both the vocabulary items in a word list and their actual frequencies would differ substantially if actual semantics were a consideration.

General Summary

The results of this study seem to suggest that the differences between a semantically-based and form-based frequency list are substantial. The overall findings for the 46 words investigated would change the list considerably. In turn, such differences would greatly affect word coverage estimates of the list overall as well as across written and spoken registers.

CHAPTER FIVE

Discussion and Conclusion

The current study is a response to a call for new, more valid high frequency word lists, particularly those intended for pedagogical purposes (Read, 2000; Gardner, 2007) and focuses on the existence of homography associated with word forms. The study analyzed 46 randomly selected, high frequency lemmas to investigate how extensive the existence of homography is for such lexical forms in a representative sample of English (the BNC), by determining possible lexemes for each form and calculating total occurrences of each lexeme within each sample. A comparison between written and spoken frequencies was also performed in order to assess potential form-meaning variations between these two major subregisters.

The results indicate that high frequency word forms demonstrate a considerable amount of homography (cf. Ravin & Leacock, 2000) in a representative corpus of English. A discussion of the findings from this research will help to highlight the form-meaning gap in the construct of *word*, specifically with regard to computer-generated word lists in the fields of applied corpus linguistics specifically and more traditional corpus linguistics in general. The implications of this research, particularly for teaching and learning ESL, will be discussed, followed by a listing of limitations of the study and suggestions for future research.

To begin, only one other study (Ming-tzu & Nation, 2004) has looked at the effects of homography on word lists. That study specifically investigated homography in the Academic Word List (AWL) and showed that roughly 10% of the word families in the AWL were affected by significant homography. By contrast, the results of the

present study suggest a much higher incidence (50%) of homography. Of the 46 lemmas analyzed in this study, 34 demonstrated some homography, with exactly half (23) demonstrating substantial homography (meeting the 10% representation threshold). The distinct homographs (lexemes) represented in these 23 lemmas would create 35 additional entries in a high frequency word list, leading not only to a considerable change in the number of items in the list, but also in the ranking of the individual words in that list. There are several possible reasons for the discrepancies between the results of the two studies.

First, the AWL analyzed by Ming-tzu and Nation (2004) represents a set of sub-technical words from a more restricted corpus of adult academic materials (Coxhead, 2000); whereas the present study looked specifically at general high frequency words from a much broader mega-corpus (the BNC). That is, the words in the AWL are by definition more specialized than the words in the current study, and therefore less likely to exhibit homography.

A second possible reason may be how levels of semantic relatedness were determined. Both studies used the semantic relatedness scale from Nagy and Anderson's (1984) paper to determine the boundary between a polyseme and a homograph. However, Ming-tzu and Nation (2004) decided that any meanings related at levels 0–3 would be polysemous, while any relationship beyond that would distinguish meanings as separate homographs. By contrast, this study set a slightly more conservative standard, cutting polysemy off at level 2, primarily based on the fact that ESL learners are known to struggle more with the transparency of form-meaning relationships and the semantic boundaries of words than native speakers (Al-Ali, 2004; Jiang, 2004; Nation, 2001b;

Zughoul, 1991). The cutoff at level 2 also mirrors the threshold established in the original Nagy and Anderson (1984) study to distinguish semantic transparency (0-2) from semantic opacity (3-5).

Perhaps the most crucial finding of this study is that the construct used to define a word (*word family* vs. *lemma* vs. *lexeme*) and to count its frequency is vitally important in creating a valid high frequency English word list. With regard to the results of this study, where word forms with their actual meanings (lexemes) are contrasted with word forms only (lemmas), there is no question that the choice of construct will have a marked impact on the size of the list (potentially many more items if lexemes are counted) and the rankings of individual words on those lists (forms with multiple meanings—lexemes—would either be reduced in rank on the list, raised in rank, or would leave the list altogether). In fact, if the findings for the 46-item analysis were consistent for the remainder of the high frequency forms, a potential high frequency list would almost double in size (i.e., 35 new items added to the 46-item list). This is by no means a minor issue when it comes to both the pedagogical and research purposes for such a list.

The impact of homography found in this study also supports the cautionary notes made by several researchers regarding the form-meaning disconnect in computer-generated frequency counts and word lists (i.e. Engels, 1968; Gardner, 2007; Read, 2000). The findings also validate concerns enumerated by Gardner (2007) about the psychological validity of using the "lemma' and the even more liberal "word family" to operationalize the construct of *word*, as well as Read's (2000) general call for a new high frequency word list, based on these and other important considerations.

The two additional research questions of this study were designed to focus the attention of the findings to more practical applications in TESOL and Applied Linguistics.

Research Question 2a: What are the implications of these lexical findings for estimates of vocabulary coverage in texts (written and spoken)?

One crucial implication for the findings of this study is that the construct used to define a word and to count its frequency is vitally important in making accurate estimates of the number of "words" in written and spoken English texts. Simply put, the fact that consistent forms often have multiple meanings among the high frequency words of English suggests that the lexical composition of written and spoken texts is much more complex than traditional corpus-based estimates have indicated. Additionally, the findings in this study also point to differences in the form-meaning relationships across spoken and written registers, again suggesting that overall lexical complexity in the language may have been underestimated in many studies found in the TESOL and Applied Linguistics literature.

An example of the effects of such oversights would be with concordancing, a popular corpus-based tool that is often suggested as a way to expose ESL learners quickly to target words in numerous contextualized scenarios from authentic materials—the key being that these examples come one after the other on the concordancing screen (key words in context--KWIC). If these words are higher frequency and exhibit homography like many of the words in the current study, it is clear that this practice could place a burden on language learners trying to disambiguate one meaning from another in context. This would be especially true if the sentences containing the words were drawn randomly

from the electronic text and if no distinction were made between similar word forms drawn from different registers (e.g., written vs. spoken English).

Research Question 2b: What are the implications of these lexical findings for the teaching and learning burden of vocabulary in ESL contexts?

By extension of the discussion above, traditional coverage estimates of computer-generated lists (based on word families or lemmas) have been high, while traditional estimates of the number of running words that L2 learners must know in order to have "adequate comprehension" (Nation 2006, p. 61) have been low. Both of these distortions have important pedagogical ramifications with regard to the use of word lists for pedagogical purposes. For one, research suggests that L2 learners must know 95-98% of the running words in the text for basic comprehension to take place (Nation 2006). Until recently, word-family advocates have suggested that knowledge of roughly 2,500 to 3,000 high frequency word families would allow L2 learners to reach the 95% threshold—a view that continues to be espoused by many in the field. Thus, directly teaching the words on these lists was thought to be both essential and feasible, given the relatively low number of items (2,500-3,000).

However, results of the current study suggest that such form-based lists do not accurately reflect the true nature of high-frequency word forms, which often have multiple meanings, thus posing an instructional burden as teachers are left to decide which of the many potential meanings to teach. Additionally, when these lists are used to assess the lexical composition of ESL materials, they will tend to underestimate the number of vocabulary items and the relative lexical density of the materials, as well as the vocabulary knowledge necessary to negotiate the meaning of the materials. The form-

meaning disparities noted between written and spoken English also suggest that ESL teachers may need to teach the multiple meanings of homonymous forms in order for learners to be adequately prepared to negotiate the basic meanings in both registers. However, some research suggests that teachers should not introduce multiple meanings of the same word form at the same time (Folse, 2004).

With regard to research involving ESL pedagogy and language acquisition, the findings of this study strongly caution that the construct of *word* needs to be more carefully scrutinized, particularly in using corpora and computers to generate word lists for teaching, measuring coverage, or assessing requisite vocabulary knowledge. Essentially, when large form-based word families or smaller form-based lemma groupings are used to assess coverage, two premises are often assumed: 1) that the ESL students will be able to make the same word family or lemma connections as native speakers, and 2) that ESL students will know the various homographs associated with the forms included in word-family groupings (all inflectionally- and derivationally-related forms) or lemma groupings (all inflectionally-related forms). However, there is ample evidence to suggest that these notions are false (Bauer & Nation, 1993; Coniam, 1999; Nation 2006; Stubbs, 2002).

Given the extent of homography noted in this study, the second premise is particularly troublesome. A brief discussion of this may therefore be warranted. To begin, the assumption of form-meaning transparency within word families may be true some of the time. For example, if one knows the English word *light,* the meanings of the separate lemmas LIGHT (n), LIGHT (v), and LIGHT (adj) may seem obvious. However, in the ESL context, one must be careful to assume such connections, since the L1 translations of

a given word may represent very distinct lexemes (forms and meaning are different) from those in English (forms similar, but meanings different). For example, in Spanish the lemma LIGHT (n) is ***luz***, but the Spanish word for the lemma LIGHT (v) in ***light*** *a candle* is ***encender*** and in *she* ***lights*** *the room with her presence* is ***iluminar*** or ***alumbrar*****.** The even more obscure verb, as in *the bird* ***lights*** *on the branch,* is ***posarse***. This problem is exacerbated even more as the existence of homography among the English meanings is discovered. For example, in English one says *he wore a* ***light*** *jacket,* which in Spanish is *el se vistio de una chaqueta* ***ligera,*** and when talking about colors, like ***light*** *green* in Spanish, it is said *verde* ***claro***. Additionally, the more obscure use of *light* in English, (i.e.--*there was a* ***light*** *breeze*) is expressed in Spanish as *una brisa* ***suave***. In analyzing just one of many English examples like *light*, it is easy to see that in some languages the psychological relatedness of lexemes and lemmas that share the same form would not necessarily be transparent and may need to be taught. This has also been pointed out in other research using different examples (Nation 2001b).

The example of the homograph *light* indicates that native Spanish speakers would likely need to reconfigure the semantic boundaries of *light* when working with both their L1 background and the L2 word. As illustrated above, differences in semantic boundaries affect the learning burdens of words in L2 acquisition. Certainly, this burden intensifies when homography exists because additional distinct meanings attached to a single word form indubitably lead to more potential confusion and require greater linguistic knowledge for disambiguation of senses and eventual comprehension and production (e.g., *father* – a priest; *father* – a parent; *godfather*; to *father a child*). Nation (2001b) points this out when he states that “the strength of the connection between the

form and its meaning will determine how readily the learner can retrieve the meaning when seeing or hearing the word form, or retrieve the word form when wishing to express the meaning" (p. 48).

From a pedagogical perspective, the finding in this study that 50% of the high frequency lemma forms had only one meaning is also noteworthy. From both a teacher and a learner perspective, knowing that roughly half of the high frequency word forms have only one meaning is beneficial in approaching the task of vocabulary teaching and learning, provided that the these non-homonymous words are known. ESL teachers and learners would specifically want to target such words, particularly in initial stages of learning because they are very productive and more straightforward to teach, learn, and use.

Limitations

The following limitations became evident during the course of the study:

1. This study was obviously limited by the sample size, which only represented about 1% of the 4,277 lemmas being defined as high frequency. However, there is no reason to believe that the findings would vary differently if a larger sample had been used, especially given the random sampling procedures that were employed. The reduction to 46 from the original 100 lemmas was a result of time and resource constraints, which attest to the difficulty of performing manual semantic tagging

2. Only one sample of English was used—the BNC:

- British English may exclude lexical items and meanings that may be more frequent in American English.
- The BNC, published in 1993, may be slightly outdated.

- The BNC has a disproportionately small representation of spoken language, which may have skewed the actual percentages of sense representation found in this study.

3. Some subjectivity was involved in determining levels of semantic relatedness, though a consensus was reached through discussions between raters.

4. Multi-word items or phraseology of any sort was excluded even though these lexical items are recurrent throughout the language and qualify as unique lexical items. Often, they were not included in the sense possibilities and therefore posed a problem for raters who frequently encountered them and had to deal with them on an individual basis, instead of systematically rating them.

5. Many nouns or adjectives were part of compound nouns that were sometimes loosely and sometimes tightly bound. The word SCHOOL (n) was very frequently subject to this, for example, *school uniform, school age, school teacher, school system, school children* and *school districts*. Each decision was made on an individual basis at various times throughout the rating process.

6. There was no specific method in the selection of raters and specific inter-rater reliability was not performed, although triple ratings were often used to determine sense ratings.

Suggestions for further research

The most obvious suggestion for future research is for linguists and researchers to come to a consensus on how to define the construct of word in a psychologically-valid way, with particular attention to meaning as well as form-based relationships.

An obvious subsequent step in the research is the development of a semantically-based high frequency list. As Read (2000) suggests, a good list would take into consideration the following: 1) a core or basic vocabulary, derived from accurate lemma groupings, and tailored somewhat to a general usage of English (like the GSL but more current) and supplemented (from a pedagogical standpoint) by additional words more specifically chosen for teaching and learning goals, and 2) inclusion of multi-word units, such as phrasal verbs, tightly bound compound nouns, stock phrases, high frequency idioms, and so forth.

In addition to a core vocabulary, other high frequency lists tailored to English for various specific purposes could be created. To make such lists feasible, corpus linguists should continue to work on computer programs that are able to identify lexemes. This may be more realistic with smaller, more distinct registers to begin with.

With regard to instruction and learning, it is important to create lists based on theoretically and methodologically sound principles. Corpora rich with linguistic information can be useful in developing many lists from which effective teaching material can be produced. Creation of a list should also be tailored to the purpose for the list (written vs. spoken English, beginning vs. advanced learners, curriculum specific lists, etc.). Semantically-valid lists could also be useful in creating graded readers, simplifying texts, and improving assessment and testing materials. Finally, caution must always be exercised in using computer-generated information for teaching and learning purposes.

References

Al-Ali, M. (2004). Familiar words in unfamiliar contexts. *Perspectives: Studies in translatology, 12*(2), 134-144.

Anderson, R.C. & Nagy, W.E. (1991). Word meanings. In R. Barr *et al's Handbook of Reading Research*, Volume II (pp. 690-794). White Plains, NY: Longman Publishing Group.

Bauer, L. & Nation, P. (1993). Word families. *International Journal of Lexicography, 6* (4), 253-277.

British National Corpus. Retrieved on November 19, 2008 from http://www.natcorp.ox.ac.uk/.

Carter, R. (1998). *Vocabulary: Applied linguistic perspectives, 2nd ed*. New York: Routledge.

Coniam, D. (1999). An investigation into the use of word frequency lists in computing vocabulary profiles. *Hong Kong Journal of Applied Linguistics, 4*(1), 103-122.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Darwin, C.M., & Gray, L.S. (1999). Going after the phrasal verb: An alternative approach to classification. *TESOL Quarterly, 33(1),* 65-83.

Davies, M. (2008). Retrieved on November 19, 2008 from http://corpus.byu.edu/bnc/.

Engels, L.K. (1968). The fallacy of word counts. *International Review of Applied Linguistics in Language Teaching, 6(3),* 213-231.

Folse, K.S. (2004). *Vocabulary myths: Applying second language research to classroom teaching.* Ann Arbor: The University of Michigan Press.

Gardner, D. (2007). Validating the Construct of *Word* in Applied Corpus-based Vocabulary Research: A Critical Survey. *Applied Linguistics, 28(2),* 241-265.

Haastrup, K. & Henriksen, B. (2000). Vocabulary acquisition: Acquiring depth of knowledge through network building. *International Journal of Applied Linguistics, 10(2),* 221-239.

Harris, A. J., & Jacobson, M. D. (1973-1974). Comparing word lists. *Research Reading Quarterly 1(IX/1),* 87-109.

Heatley, A., Nation, I.S.P., & Coxhead, A. (2002). RANGE and FREQUENCY programs. Retrieved November 17, 2008 from http://www.vuw.ac.nz/lals/staff/Paul_Nation.

Ijaz, I.H. (1986). Linguistic and cognitive determinants of lexical acquisition in second language. *Language Learning, 36,* 401-451.

Jiang, I. (2004). Semantic transfer and its implications for vocabulary teaching in a second language. *The Modern Language Journal 88(3),* 416-432.

Kilgarrif, A. (2006). Retrieved on October 1, 2007 from http://www.kilgarriff.co.uk/bnc-readme.html.

Knowles, G. & Don, Z.M.. (2004). The notion of a "lemma": Headwords, roots and lexical sets. *International Journal of Corpus Linguistics. 9(1*), 69-81.

Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *Canadian Modern Language Review, 59(4),* 567-587.

Miller, G.A. (1999). On knowing a word. *Annual Review of Psychology, 50,* 1-19.

Miller, G.A. (2006). Retrieved on November 19, 2008 from http://wordnet.princeton.edu.

Ming-Tzu, K.W. & Nation, P. (2004). Homography in the academic word list. *Applied Linguistics, 25*(3), 291-314.

Nagy, W. E. & Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly, 19(3)*, 304-330.

Nagy, W.E., & Herman, P.A. (1987). Breadth and depth of vocabulary knowledge: implications for acquisition and instruction. In M.G. Mckeown and M.E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 19-35). Hillsdale, NJ: Lawrence Erlbaum Associates.

Nation, I.S.P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review. 63*(1), 59-82.

Nation, I.S.P. (2001a). How many high frequency words are there in English? From M. Gill, A.W. Johnson, L. M. Kiski, R.D. Sell, & B. Warvik (Eds.). *Language, learning, literature: Studies presented to Hakan Kingdom* (English Department Publications 4). Turku: Abo Akademi University, 167-181.

Nation, I.S.P. (2001b). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nerlich, B., Todd, V., Herman, V., & Clarke, D.D. (Eds.). (2003). *Trends in linguistics: Polysemy – Flexible patterns of meaning in mind and language.* New York: Mouton de Grutyer.

Ogden, C.K. (1942). *The general basic English dictionary*. New York: W. W. Norton & Company, Inc.

Ogden, C.K. (1934). *The system of basic English*. New Jersey: Quinn and Boden Company, Inc.

Preller, A. G. (1967). Some problems involved in compiling word frequency lists. *The Modern Language Journal, 52*,(7), 399 – 402.

Pulido, D. (2003). Modeling the role of second language proficiency and topic familiarity in second language incidental vocabulary acquisition through reading. *Language Learning, 53*(2), 233-284.

Ravin, Y. & Leacock, C. (2000). *Polysemy: Theoretical and computational approaches.* Oxford: Oxford University Press.

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

Richards, R.C. (1974). Word lists: problems and prospects. *RELC Journal, 5*(2), 69-84.

Schmitt, N. & Zimmerman, C.B. (2002). Derivative word forms: What do learners know? *TESOL Quarterly, 36*(2), 145-171.

Sinclair, J. (2004). *Trust the text: Language, corpus and discourse.* New York: Routledge.

Stubbs, M. (2002). *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell Publishing.

Thorndike, E.L & Lorge, I. (1944). *The teacher's word book of 30,000 words.* New York: Teachers College, Columbia University.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics, 22*(2), 217-234.

Wei, M. & Light, T. (1973). *A newspaper vocabulary*. Hong Kong: The Chinese University of Hong Kong.

West, M. (1953). *A general service list of English words*. London: Longman Group LTD.

Zughoul, M.R. (1991). Lexical choice: Toward writing a problematic word list. *International Review of Applied Linguistics in Language Teaching*, *29(1)*, 46-60.

APPENDIX A

Sense Distributions: significant homography list and little or no homography list

SIGNIFICANT LEVELS OF HOMGRAPHY

Sense Distributions – each number is the % of the total that each sense represents

	LEMMA + POS	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12	s13	s14	s15	s16
1	WORK (v)	27	31	20	2	3	1	0	0	1	1	14	0	1	0	1	1
2	CHARACTER (n)	32	32	13	14	9	2	--	--	--	--	--	--	--	--	--	--
3	BUSINESS (n)	32	35	8	24	1	0	1	--	--	--	--	--	--	--	--	--
4	MEMORY (n)	36	30	35	--	--	--	--	--	--	--	--	--	--	--	--	--
5	ACT (v)	41	41	18	0	--	--	--	--	--	--	--	--	--	--	--	--
6	BATH (n)	6	47	35	0	1	11	--	--	--	--	--	--	--	--	--	--
7	DIRECTION (n)	47	43	6	3	--	--	--	--	--	--	--	--	--	--	--	--
8	APPLICATION (n)	24	49	3	24	1	--	--	--	--	--	--	--	--	--	--	--
9	STAND (v)	50	6	13	7	8	7	0	4	1	3	1	3	--	--	--	--
10	DEVELOP (v)	51	47	1	1	0	--	--	--	--	--	--	--	--	--	--	--
11	WELL (adj)	45	51	4	--	--	--	--	--	--	--	--	--	--	--	--	--
12	ANSWER (n)	53	47	--	--	--	--	--	--	--	--	--	--	--	--	--	--
13	GAS (n)	29	12	0	1	62	--	--	--	--	--	--	--	--	--	--	--
14	SUBJECT (n)	63	21	1	12	1	1	1	--	--	--	--	--	--	--	--	--
15	FAIR (adj)	68	23	4	0	2	0	1	2	--	--	--	--	--	--	--	--
16	SHOW (v)	69	28	2	0	0	1	1	--	--	--	--	--	--	--	--	--
17	AWFUL (adj)	70	30	0	--	--	--	--	--	--	--	--	--	--	--	--	--
18	GREEN (adj)	71	10	2	1	1	2	12	1	--	--	--	--	--	--	--	--
19	PULL (v)	72	4	6	1	0	0	15	1	1	1	1	--	--	--	--	--
20	BUSY (adj)	75	25	1	--	--	--	--	--	--	--	--	--	--	--	--	--
21	BACK (adv)	78	15	7	--	--	--	--	--	--	--	--	--	--	--	--	--
22	MATCH (n)	11	81	4	0	2	3	1	0	--	--	--	--	--	--	--	--
23	REST (n)	87	12	1	0	0	--	--	--	--	--	--	--	--	--	--	--

This table shows the 23 Lemmas that had two or more senses that occurred over 10% of the time. The chart starts with the lemmas with the greatest degree of homography and descends through the last lemma that falls in this category.

LITTLE OR NO HOMOGRAPHY

Sense Distributions – each number is the % of the total that each sense represents

	LEMMA + POS	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12	s13	s14	s15	s16
1	DIFFICULT (adj)	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
2	NECESSARY (adj)	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
3	PRESUMABLY (adv)	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
4	PROPERLY (adv)	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
5	UNIVERSITY (n)	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
6	WHILE (n)	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
7	RESPONSIBILITY (n)	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
8	CONGRESS (n)	100	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--
9	POLICE (n)	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
10	CLOTHES (n)	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
11	MIDDLE (n)	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
12	TRAIN (v)	100	0	0	0	0	--	--	--	--	--	--	--	--	--	--	--
13	MARRY(v)	99	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--
14	MUM (n)	0	99	1	--	--	--	--	--	--	--	--	--	--	--	--	--
15	YESTERDAY (adv)	98	2	--	--	--	--	--	--	--	--	--	--	--	--	--	--
16	START (n)	98	1	1	--	--	--	--	--	--	--	--	--	--	--	--	--
17	SCHOOL (n)	97	3	0	--	--	--	--	--	--	--	--	--	--	--	--	--
18	PAGE (n)	97	0	0	2	--	--	--	--	--	--	--	--	--	--	--	--
19	NORTH (n)	1	97	0	2	--	--	--	--	--	--	--	--	--	--	--	--
20	SLEEP (v)	97	1	3	--	--	--	--	--	--	--	--	--	--	--	--	--
21	TRY (v)	92	7	1	0	1	0	0	0	--	--	--	--	--	--	--	--
22	ROAD (n)	91	6	--	--	--	--	--	--	--	--	--	--	--	--	--	--
23	NAME (n)	88	4	5	3	1	--	--	--	--	--	--	--	--	--	--	--

This table shows the 23 Lemmas that had one sense that occurred over 90% of the time (or two senses did not occur 10% of the time or more). The chart starts with the lemmas with the lowest degree of homonomy and end with the last lemma that falls into the described category. Numbers represent the percent coverage of each sense.

Appendix B

Complete lemmatized frequency list of the 46 lemmas and GSL rankings
and
Complete lexeme-based frequency list of the 81 lexemes

Lemmatized frequency list (based on Kilgarriff's results from the BNC)

	Lemmas	Lemmatized ranking	Lemmatized frequency count	GSL Ranking
1	BACK (adv)	118	75,494	90
2	WORK (v)	129	67,842	71
3	SHOW (v)	163	58,152	123
4	TRY (v)	174	54,422	222
5	SCHOOL (n)	181	52,227	137
6	BUSINESS (n)	244	38,204	265
7	STAND (v)	285	32,899	194
8	NAME (n)	292	32,309	225
9	SUBJECT (n)	336	29,091	434
10	POLICE (n)	360	27,508	679
11	DEVELOP (v)	411	24,205	130
12	ROAD (n)	441	23,103	440
13	DIFFICULT (adj)	466	22,033	674
14	YESTERDAY (adv)	533	19,459	1033
15	NECESSARY (adj)	573	18,107	387
16	APPLICATION (n)	623	16,281	1005
17	ACT (v)	654	15,620	273
18	PAGE (n)	708	14,546	952
19	REST (n)	712	14,440	465
20	PULL (v)	743	13,852	685
21	ANSWER (n)	810	12,596	412
22	CHARACTER (n)	818	12,511	559
23	RESPONSIBILITY (n)	846	12,078	Responsible = 514 (same word family)
24	TRAIN (v)	855	11,907	363
25	UNIVERSITY (n)	893	11,367	448
26	DIRECTION (n)	924	10,905	634
27	MEMORY (n)	984	10,221	827
28	START (n)	1,083	9,268	246
29	GREEN (adj)	1,116	9,013	872
30	NORTH (n)	1,123	8,949	551
31	MATCH (n)	1,159	8,718	927
32	MARRY (v)	1,171	8,631	777
33	GAS (n)	1,226	8,133	895
34	CLOTHES (n)	1,331	7,308	Clothe = 890 (same word family)
35	SLEEP (v)	1,381	6,992	749
36	FAIR (adj)	1,393	6,936	666
37	MIDDLE (n)	1,509	6,363	850

38	WHILE (n)	1,578	6,058	142
39	PROPERLY (adv)	1,680	5,667	Proper = 711 (same word family)
40	CONGRESS (n)	1,708	5,544	—
41	BUSY (adj)	1,801	5,221	1179
42	BATH (n)	2,439	3,484	1717
43	PRESUMABLY (adv)	2,532	3,279	—
44	AWFUL (adj)	2,731	2,960	—
45	WELL (adj)	3,247	2,241	96
46	MUM (n)	—	—	—

Lexeme-based frequency list (based on Kilgarriff's results with extrapolations)

	Lexemes	Lexeme-based ranking	Lexeme-based frequency count
1	[BACK- 1] (adv)	162	58,719
2	[SCHOOL] (n)*	186	50,660
3	[TRY] (v)*	189	50,068
4	[SHOW- 1] (v)	234	39,886
5	[POLICE] (n)	360	27,508
6	[NAME] (n)*	346	28,432
7	[DIFFICULT] (adj)	466	22,033
8	[WORK- 2] (v)	488	21,133
9	[ROAD] (n)*	488	21,024
10	YESTERDAY (adv)*	542	19,070
11	[SUBJECT- 1] (n)	564	18,470
12	[WORK-1] (v)	573	18,168
13	[NECESSARY] (adj)	573	18,107
14	[STAND- 1] (v)	619	16,450
15	[SHOW- 2] (v)	619	16,440
16	[PAGE] (n) *	734	14,110
17	[WORK- 3] (v)	761	13,345
18	[BUSINESS- 2] (n)	771	13,180
19	[REST- 1] (n)	809	12,616
20	[DEVELOP- 1] (v)	829	12,351
21	[BUSINESS- 1] (n)	838	12,225
22	[RESPONSIBILITY] (n)	846	12,078
23	[TRAIN] (v)	855	11,907
24	[DEVELOP- 2] (v)	885	11,478
25	[BACK- 2] (adv)	887	11,437
26	[UNIVERSITY] (n)	893	11,367
27	[PULL- 1] (v)	1009	10,008
28	[WORK-11] (v)	1046	9,640

29	[START] (n) *	1,110	9,083
30	[BUSINESS- 4] (n)	1,120	8,978
31	[NORTH] (n) *	1,161	8,681
32	[MARRY] (v) *	1,171	8,631
33	[APPLICATION- 2] (n)	1,241	7,978
34	[CLOTHES] (n)	1,331	7,308
35	[SLEEP] (v)*	1,426	6,782
36	[ANSWER- 1] (n)	1,437	6,717
37	[GREEN- 1] (adj)	1,491	6,438
38	[ACT- 1] (v)	1,494	6,432
39	[ACT- 2] (v)	1,495	6,432
40	[MIDDLE] (n)	1,509	6,363
41	[SUBJECT- 2] (n)	1,556	6,156
42	[WHILE] (n)	1,578	6,058
43	[ANSWER- 2] (n)	1,623	5,879
44	[PROPERLY] (adv)	1,680	5,667
45	[CONGRESS] (n)	1,708	5,544
46	[MATCH- 2] (n)	1,747	5,417
47	[BACK- 3] (adv)	1,773	5,337
48	[DIRECTION- 1] (n)	1,798	5,228
49	[GAS- 5] (n)	1,918	4,811
50	[DIRECTION- 2] (n)	1,942	4,722
51	[FAIR- 1] (adj)	1,943	4,715
52	[STAND- 3] (v)	2,161	4,112
53	[CHARACTER-2] (n)	2,210	4,004
54	[CHARACTER-1] (n)	2,237	3,941
55	[APPLICATION- 4] (n)	2,254	3,907
56	[BUSY- 1] (adj)	2,255	3,902
57	[APPLICATION- 1] (n)	2,277	3,826
58	[MEMORY- 1] (n)	2,360	3,632
59	[SUBJECT- 4] (n)	2,414	3,540

60	[MEMORY- 3] (n)	2,421	3,528
61	[PRESUMABLY] (adv)	2,532	3,279
62	[MEMORY- 2] (n)	2,665	3,061
63	[ACT- 3] (v)	2,854	2,757
64	[GAS- 1] (n)	3,152	2,342
65	[PULL- 7] (v)	3,382	2,103
66	[AWFUL- 1] (adj)	3,404	2,078
67	[CHARACTER- 4] (n)	3,937	1,689
68	[REST- 2] (n)	3,957	1,678
69	[BATH- 2] (n)	4,017	1,641
70	[FAIR- 2] (adj)	4,096	1,597
71	[CHARACTER-3] (n)	4,146	1,564
72	[BUSY- 2] (adj)	4,711	1,292
73	[BATH- 3] (n)	4,865	1,235
74	[WELL- 2] (adj)	5,102	1,148
75	[GREEN- 7] (adj)	5,265	1,090
76	[WELL- 1] (adj)	5,531	1,000
77	[GREEN- 2] (adj)	5,745	941
78	[GAS- 2] (n)	5,756	937
79	[AWFUL- 2] (adj)	5,943	882
80	[MATCH- 1] (n)	___	738
81	[BATH- 6] (n)	___	369
82	[MUM] (n)	___	+

*minimally modified, but only one sense represented at the criterion of 10% representation

Appendix C

Sense Distributions of Written vs. Spoken:
significant homography list and little or no homography list

SIGNIFICANT LEVELS OF HOMGRAPHY
Written vs. spoken

	LEMMA + POS		s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12	s13	s14	s15	s16
1	WORK (v)	S	24	29	20	2	1	1	0	0	1	0	20	0	0	0	1	0
		W	30	33	19	1	6	0	0	0	0	1	8	0	1	0	0	1
2	CHARACTER (n)	S	23	39	11	17	9	1	--	--	--	--	--	--	--	--	--	--
		W	40	25	14	10	9	2	--	--	--	--	--	--	--	--	--	--
3	BUSINESS (n)	S	33	26	9	30	2	0	0	--	--	--	--	--	--	--	--	--
		W	31	43	7	17	0	0	2	--	--	--	--	--	--	--	--	--
4	MEMORY (n)	S	32	27	41	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	39	33	28	--	--	--	--	--	--	--	--	--	--	--	--	--
5	ACT (v)	S	41	49	11	0	--	--	--	--	--	--	--	--	--	--	--	--
		W	41	35	23	0	--	--	--	--	--	--	--	--	--	--	--	--
6	BATH (n)	S	1	56	34	0	0	9	--	--	--	--	--	--	--	--	--	--
		W	11	39	36	0	1	12	--	--	--	--	--	--	--	--	--	--
7	DIRECTION (n)	S	49	47	3	1	--	--	--	--	--	--	--	--	--	--	--	--
		W	47	40	9	4	--	--	--	--	--	--	--	--	--	--	--	--
8	APPLICATION (n)	S	22	61	2	15	0	--	--	--	--	--	--	--	--	--	--	--
		W	25	37	3	33	2	--	--	--	--	--	--	--	--	--	--	--
9	STAND (v)	S	49	5	9	6	11	9	0	5	1	2	1	2	--	--	--	--
		W	51	6	16	7	5	5	0	3	0	3	0	3	--	--	--	--
10	DEVELOP (v)	S	42	55	2	1	0	--	--	--	--	--	--	--	--	--	--	--
		W	60	40	0	0	0	--	--	--	--	--	--	--	--	--	--	--
11	WELL (adj)	S	50	40	10	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	41	59	0	--	--	--	--	--	--	--	--	--	--	--	--	--
12	ANSWER (n)	S	38	63	--	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	69	31	--	--	--	--	--	--	--	--	--	--	--	--	--	--
13	GAS (n)	S	16	12	0	0	72	--	--	--	--	--	--	--	--	--	--	--
		W	40	11	0	1	47	--	--	--	--	--	--	--	--	--	--	--
14	SUBJECT (n)	S	75	16	1	4	1	1	1	--	--	--	--	--	--	--	--	--
		W	53	26	1	19	1	0	0	--	--	--	--	--	--	--	--	--
15	FAIR (adj)	S	69	27	0	0	2	0	1	0	--	--	--	--	--	--	--	--
		W	67	19	8	0	2	0	0	4	--	--	--	--	--	--	--	--
16	SHOW (v)	S	70	25	2	0	0	1	2	--	--	--	--	--	--	--	--	--
		W	67	32	1	0	0	0	0	--	--	--	--	--	--	--	--	--

17	AWFUL (adj)	S	60	40	0	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	80	20	0	--	--	--	--	--	--	--	--	--	--	--	--	--
18	GREEN (adj)	S	66	13	1	1	0	2	16	1	--	--	--	--	--	--	--	--
		W	78	8	3	1	1	1	8	0	--	--	--	--	--	--	--	--
19	PULL (v)	S	71	3	5	2	0	0	18	0	1	0	0	--	--	--	--	--
		W	73	4	6	0	0	0	13	1	0	1	1	--	--	--	--	--
20	BUSY (adj)	S	84	15	1	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	66	34	0	--	--	--	--	--	--	--	--	--	--	--	--	--
21	BACK (adv)	S	88	7	5	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	68	23	9	--	--	--	--	--	--	--	--	--	--	--	--	--
22	MATCH (n)	S	16	79	3	0	0	1	0	0	--	--	--	--	--	--	--	--
		W	6	82	4	0	3	4	1	0	--	--	--	--	--	--	--	--
23	REST (n)	S	94	6	0	0	0	--	--	--	--	--	--	--	--	--	--	--
		W	81	17	2	0	0	--	--	--	--	--	--	--	--	--	--	--

This table shows the 23 lemmas with significant homography and compares sense representation in the written and spoken registers. Numbers represent the percent coverage of each sense.

LITTLE OR NO HOMOGRAPHY

Written vs. spoken

LEMMA + POS		s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12	s13	s14	s15	s16	
1	DIFFICULT (adj)	S	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
2	NECESSARY (adj)	S	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
3	PRESUMABLY (adv)	S	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
4	PROPERLY (adv)	S	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
5	UNIVERSITY (n)	S	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
6	WHILE (n)	S	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
7	RESPONSIBILITY (n)	S	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
8	CONGRESS (n)	S	100	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	100	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--
9	POLICE (n)	S	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
10	CLOTHES (n)	S	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
11	MIDDLE (n)	S	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	100	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
12	TRAIN (v)	S	100	0	0	0	0	--	--	--	--	--	--	--	--	--	--	--

		W	100	0	0	0	0	--	--	--	--	--	--	--	--	--	--	--
13	MARRY (v)	S	100	0	--	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	99	1	--	--	--	--	--	--	--	--	--	--	--	--	--	--
14	MUM (n)	S	0	99	1	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	0	99	1	--	--	--	--	--	--	--	--	--	--	--	--	--
15	YESTERDAY (adv)	S	98	2	--	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	98	2	--	--	--	--	--	--	--	--	--	--	--	--	--	--
16	START (n)	S	99	1	0	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	97	1	2	--	--	--	--	--	--	--	--	--	--	--	--	--
17	SCHOOL (n)	S	99	1	0	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	95	5	0	--	--	--	--	--	--	--	--	--	--	--	--	--
18	PAGE (n)	S	99	0	0	1	--	--	--	--	--	--	--	--	--	--	--	--
		W	95	0	0	2	--	--	--	--	--	--	--	--	--	--	--	--
19	NORTH (n)	S	0	100	0	0	--	--	--	--	--	--	--	--	--	--	--	--
		W	2	94	0	4	--	--	--	--	--	--	--	--	--	--	--	--
20	SLEEP (v)	S	99	0	1	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	95	1	4	--	--	--	--	--	--	--	--	--	--	--	--	--
21	TRY (v)	S	89	10	0	0	1	0	0	0	--	--	--	--	--	--	--	--
		W	95	4	1	0	0	0	0	0	--	--	--	--	--	--	--	--
22	ROAD (n)	S	92	2	--	--	--	--	--	--	--	--	--	--	--	--	--	--
		W	90	10	--	--	--	--	--	--	--	--	--	--	--	--	--	--
23	NAME (n)	S	96	2	1	1	0	--	--	--	--	--	--	--	--	--	--	--
		W	81	6	8	4	1	--	--	--	--	--	--	--	--	--	--	--

Numbers represent the percent coverage of each sense.

Appendix D

Contrast between lemmatized and form frequency counts

Lemma	Lemmatized Freq. Count	Form Freq. Count	POS Included in Form Freq. Count
NAME (n)	32309	38593	noun, verb
WORK (v)	67842	130090	noun, verb
SCHOOL (n)	52227	52227	noun
WELL (adj)	2241	145852	adjective, adverb, interjection, noun
GREEN (adj)	9013	12183	adjective, noun
DIFFICULT (adj)	22033	22033	adjective
SUBJECT (n)	29091	32392	noun, adjective, verb
TRY (v)	54422	55799	verb, noun
MARRY (v)	8631	8631	verb
UNIVERSITY (n)	11367	11367	noun
YESTERDAY (adv)	19459	19459	adverb
BUSY (adj)	5221	5221	adjective
WHILE (n)	6058	56606	noun, conjunction
MUM (n)	Not listed	Not listed	noun, adjective
ROAD (n)	23103	23103	noun
RESPONSIBILITY (n)	12078	12078	noun
START (n)	9268	50297	noun, verb
CONGRESS (n)	5544	5544	noun
MATCH (n)	8718	14626	noun, verb
BACK (adv)	75494	106315	adverb, adjective, noun, verb
GAS (n)	8133	8133	noun
FAIR (adj)	6936	9635	adjective, adverb, noun
PRESUMABLY (adv)	3279	3279	adverb
REST (n)	14440	19146	noun, verb
SLEEP (v)	6992	10624	verb, noun
PAGE (n)	14546	14546	noun
AWFUL (adj)	2960	2960	adjective
NORTH (n)	8949	8949	noun
POLICE (n)	27508	27508	noun
NECESSARY (adj)	18107	18107	adjective
PROPERLY (adv)	5667	5667	adverb
TRAIN (v)	11907	20127	verb, noun
MEMORY (n)	10221	10221	noun
STAND (v)	32899	37303	verb, noun
DIRECTION (n)	10905	10905	noun
CLOTHES (n)	7308	7308	noun
BUSINESS (n)	38204	38204	noun
CHARACTER (n)	12511	12511	noun
MIDDLE (n)	6363	10850	noun, adjective
ANSWER (n)	12596	22736	noun, verb
SHOW (v)	58152	70231	verb, noun
PULL (v)	13852	13852	verb
BATH (n)	3484	3484	noun
ACT (v)	15620	38277	verb, noun
APPLICATION (n)	16281	16281	noun
DEVELOP (v)	24205	24205	verb

-See Appendix D for a complete table of the 46 lemmas

-See Table 3 or Appendix B for examples of a lexeme-based frequency count