WHAT ARE THE COLLOCATIONAL EXEMPLARS OF HIGH-FREQUENCY ENGLISH VOCABULARY? ON IDENTIFYING MWUS MOST REPRESENTATIVE OF HIGH-FREQUENCY, LEMMATIZED CONCGRAMS

A thesis submitted by

James M. Rogers, M TESOL

for the award of

Doctor of Philosophy

2016

Abstract

Collocations, simply defined, are words that have a high frequency of co-occurrence (Biber et al., 1999: Shin, 2006). Collocational fluency is an essential aspect of communicating in and comprehending a second language in a native-like fashion. However, second language learners of English struggle to obtain such fluency since there is a lack of focus on it in the classroom and in ESL resources. This stems from the lack of a large-scale resource that identifies which collocations to teach to help learners master high-frequency English. So, although a large number of researchers agree upon the importance of collocational fluency and focusing on high-frequency collocations directly, learners, teachers and materials writers lack guidance as to which items to focus on.

Such a resource is not available because research that has consideration for all the important aspects of identifying collocations that previous researchers have identified has yet to be implemented on a large scale. Therefore, this thesis set out to accomplish such a task. The goal was to create a methodology which would result in a practical resource which identifies multi-word units most representative of high-frequency collocations of high-frequency lemma of English, and which of these items would be most useful for Japanese learners to study. It aimed to identify such items by collecting and analyzing corpus data with the help of eight native English speaking university teachers in Japan who teach English as a second language, two native English speaking junior high school teachers in Japan who teach English as a second language, five native Japanese translators with native-like ability in English, one native English speaking university professor who teaches English as a second language and has extensive knowledge developing concordance software, and one Romanian translator with native-like ability in both English and Japanese. Once identified, Japanese university freshmen were tested on their knowledge of these items.

This study took a corpus linguistics approach, working with data from the Corpus of Contemporary American English (COCA), to identify high-frequency collocations and the multi-word units they most commonly occur in. A frequency cut-off was identified which resulted in approximately 11,000 multi-word units that only consist of approximately 3,000 word families, of which the vast majority are high-frequency. Corpus dispersion and chronological data were

deemed unreliable for determining whether or not items selected had general usefulness over a variety of genres and throughout time, and time-consuming manual analysis for general usefulness was deemed essential. This was due to the fact that this study's data analysis alone would either lead to items deemed worthy of direct instruction by native speakers being flagged as having unbalanced data dispersion at certain parameters, while at other parameters items deemed unworthy of direct instruction were shown to have balanced data dispersion. Also, consideration for colligation was found to only improve upon a small percentage of items, and while useful for improving the quality of data, the process was found to be extremely complex and time consuming due to the lack of an established methodology and dedicated software. Expanding multi-word units beyond their core was found to be an essential step in that native speakers opted to do this in over half of the items identified. For example, concordance data identified *equal access* as the most frequent multi-word unit that the two lemma *equal/access* occur in (the core unit), but the native speaker opted to add the next most common multi-word unit instead (*equal access to*) in regards to what unit should be studied directly by learners. Semantic transparency analysis to help select only items that are semantically opaque and thus deserve more study time was not fruitful since the majority of items identified were considered to be semantically transparent. In contrast, L1-L2 congruency was found to be a very important criterion to consider with half of the items identified being considered incongruent to an extent, thus deserving more study time. Furthermore, native speaker intuition was found to be extremely reliable in regards to context creation using mostly high-frequency vocabulary. Out of 130,000 tokens of example sentence context created, the added content only reduced the percentage of tokens in the high-frequency realm (3,000 word families) by 0.92 percent. Confirming this was essential in that if their intuition could be relied upon for context creation that used mostly high-frequency vocabulary it would help avoid adding additional learning burden. Finally, university students' knowledge of a balanced selection of the items with consideration for frequency and L1-L2 congruency was found to be quite low overall, highlighting the need for increased focus on the list in general.

This study thus filled a major gap in the research in that it resulted in a list of items which can be utilized to help create resources or studied directly to help improve collocational fluency. A variety of steps were taken to create this resource which helped highlight the value or lack

thereof of each of these steps to achieve this study's goal. Therefore, this study should be considered a valuable contribution towards research which aims to help second language learners achieve collocational fluency.

Certification

This thesis is entirely the work of James Martin Rogers except where otherwise acknowledged. The work is original and has not previously been submitted for any other award, except where acknowledged. Student and supervisors signatures of endorsement are held at USQ.

Acknowledgements

I would like to thank and dedicate this thesis to my grandparents, my parents, my wife and daughter. Without their support none of this would have been possible.

I would also like to thank and dedicate this research to all scientists in the pursuit of knowledge and discovery, who challenge themselves to answer the most difficult questions, who do not ask self-evident questions, and who attempt to accomplish what others believe impossible. The inspiration I have received from such individuals that have come before me and the opportunity to learn from their work has proven invaluable. Without these pillars of society, we would be truly lost.

This research would also not have been possible without the helpful guidance from my main adviser Dr. Warren Midgley and second adviser Dr. Dongkwang Shin. This study would also have been difficult to accomplish if not for the technical concordance software solutions that were provided by Dr. Laurence Anthony and support and advice received from Paul Nation.

I would also like to thank all research assistants and translators that participated in this study as well. Without their help, I wouldn't have been able to tackle such a large-scale problem and provide a practical solution to it.

Table of Contents

Chapter 1

Justification of the research

Introduction

Collocations are words that have a high frequency of co-occurrence[1], and researchers agree that collocational fluency is a relatively important part of second language acquisition. However, there is evidence that students lack this knowledge, and that there is a lack of focus on developing it in the classroom. This lack of focus may be connected to a lack of a comprehensive resource that identifies which items to focus on that accurately reflects natural language. This chapter will highlight the various barriers which prevent students from obtaining collocational fluency, and thus justify the research questions that will be answered in this current study.

These research questions aim to identify the most common collocations of general English and create a resource which students can study directly and/or teacher and materials writers can use as a reference when creating ESL materials. This study does not aim to create a comprehensive resource of all collocations, but rather to identify only high-frequency items which can practically be studied or taught directly. Japanese university students' knowledge of these items will then be tested to determine the extent of their knowledge of such items.

While small collocation lists do currently exist, no large-scale list, such as what this study aims to create, has been created to date. Therefore, this study will fill a major gap in the research in the creation of such a resource and in identifying any lack of such knowledge. This study will focus on collocational knowledge among Japanese university students because this is where this thesis' author teaches English and has access to students.

Statement of the research problem

In recent years, more and more researchers are beginning to recognize the value of collocations for second language learners. Lewis (2000) stated "teaching collocation should be a top priority in every language course" (p. 8). This view stems from the realization that much of the language we speak consists of prefabricated chunks, and that collocation is one of the most

[1] Defining "collocation" is discussed in more detail later in section 2.6.1.

important kinds of chunks. Hoey (2005) and Hill (2000) also agreed that collocation plays a central role in language. So, what does this central role for collocation actually encompass for the second language learner? Multiple researchers cited how competent use of formulaic language helps the language learner to sound more natural (Durrant & Schmitt, 2009; Cowie, 1998; Wray, 2002). In addition to aiding learners in making more native-like selections, the use of collocation has been shown to make for more efficient language processing (de Glopper, 2002; Nation, 2001).

However, despite teachers being aware of the importance of collocations, their students still struggle to obtain collocational fluency (DeCock et al., 1998; Kallkvist, 1998; Waller, 1993). Research indicates that students (even advanced students) struggle with collocations, and this is a major barrier towards obtaining native-like fluency in a second language. For instance, Laufer and Waldman's (2011) finding that three separate proficiency levels of Hebrew students of English all produced far less collocations than native speakers, and that even though the amount of collocations increased at the advanced level, errors still persisted. Nesselhauf (2005) examined a 150,000 token[2] learner corpus written by advanced German learners of English, and found that a quarter of the 2,000 verb-noun collocations found were wrong, and a third deviant.

More specifically, in small scale experiments Chon and Shin (2009) and Boers et al. (2006) both found the use of formulaic expressions to correlate with perceived proficiency by native-speaker judging learners' L2 writing. Underwood, Schmitt and Galpin (2004) conducted a study examining eye-movements and found formulaic sequences to be read more quickly than non-formulaic equivalents. Conklin and Schmitt (2008) found similar results with self-paced reading tasks. Jiang and Nekrasova (2007) found grammatical judgments to be faster and more accurate for formulaic language. Regarding production, both Kuiper (1996) and Dechert (1983) found the use of formulaic language made output smoother and more fluent. Hill (2000) agrees, stating that "collocation allows us to think more quickly and communicate more efficiently" (p. 54). Furthermore, when learners utilize prefabricated language they are freeing up processing time (Almela & Sanchez, 2007; Lewis, 1993; Nation, 2001a). Furukawa et al. (1998) found that

[2] A token is every instance of a word in a corpus regardless of if the word repeats or not. For instance, in the following sentence there are 7 tokens: *This is expensive, but this is cheap.*

teaching students to utilize a chunking learning strategy improved sixth grade students' Stanford Achievement test scores by an average of 6.15 points. Sinclair (1991) referred to this in his 'idiom principle' as making "fewer and larger choices" (p. 113).

So, why is it that students lack collocational knowledge? Despite being aware of collocational fluency's importance, there is actually a severe lack of emphasis on teaching collocations. Nesselhauf's (2005) study found that the number of years learners were taught English had no positive effect on collocational knowledge. Furthermore, textbooks may not be giving students enough repetition in regards to collocation. Gitsaki's (1996) examination of a junior high school ESL textbook series found that there was very little recycling of collocations across the three books. The arbitrary nature of criteria for selecting useful collocations to teach is also contributing to the problem. The Japanese Ministry of Education (*Monbusho*, 2003) simply stated one of the goals of secondary English education is that basic collocations should be chosen for instruction, but gave no further guidelines as to which should be taught and has not updated this guideline since then.

But if teachers and materials writers are aware of the importance of collocational fluency, why not focus on them? The reason may stem from the fact that there is a severe lack of resources to refer to in selecting collocations worthwhile to study directly. Some resources do exist, but none fill the current gaps in the research.

For instance, Ackermann and Chen's (2013) *Academic Collocation List* and Simpson-Vlach and Ellis' (2010) *Academic Formulas List* only focus on academic language, while this current study aims to produce a resource for helping learners master the collocational fluency of general English. This current study also takes more advanced steps to more accurately identify co-occurrence. For example, Ackermann and Chen's resource simply lists collocations which occur adjacent to each other, while this current study takes a more advanced approach by counting co-occurrence by considering constituency variation (*lose weight* and *lose some weight* are both counted as a co-occurrence of the collocates *lose/weight*) and positional variation (*provide you support* and *support you provide* are both counted as a co-occurrence of the collocates *lose/weight*), or in other words, *concgramming*. Other resources do exist for general English, such as Martinez and Schmitt's (2012) *Phrase List*, however this list also does not consider constituency or positional variation, and is small in size (505 phrases are identified in it)

compared to this current study which in the end identified over 11,000 multi-word units (MWUs).

Many questions still remain as to which methodology produces the best results and thus this study will experiment with a variety of methods to help make it more salient how collocations and the MWUs most representative of them can be identified. Many researchers take a corpus linguistics approach to identifying high-frequency collocations. There are a variety of corpora available, but this study will utilize the Corpus of Contemporary American English (COCA) (Davies, 2008) because it is an American English corpus and data analysis and language judgments will be conducted by an American in this study. This corpus was also chosen because of its size and its balanced inclusion of a variety of language in comparison to other corpora (discussed further in section 3.3.3).

Furthermore, should collocation instruction be informed by grammatical matrices, as Mitchell (1971) and Gitsaki (1996) suggested? For example, should grammatical groups of words be counted and presented to learners, such as *[adjective] tea*, or can we limit the amount to focus on by using frequency data, grammatical well-formedness, L1 congruency, and semantic transparency, as Shin (2006) did? To elaborate, Shin utilized corpus frequency data of co-occurring words to identify high-frequency items worth studying. He also used grammatical well-formedness, or the necessity for an item identified to be a meaningful and memorizable unit. For instance, *of* and *the* co-occur often, but this does not have value for learners to study as a unit in comparison to a more meaningful and memorizable unit such as *a piece of paper*.

Shin's usage of examining the extent of L1 congruency, or the literal translation equivalents between the L2 and L1, is also an important factor to consider. For example, *eat breakfast* is literally *asagohan wo taberu* in Japanese. *Eat* literally translates into *taberu* and *breakfast* is literally *asagohan*. Thus, a Japanese learner familiar with both words and the grammatical order of English can make this MWU without a high chance of error because there is L1-L2 congruency. However, often there is not full congruency between languages. In English, we say *get credits* for a college course, but Japanese learners often make an error by literally translating how it is said in Japanese into *take credits* [tanii wo toru] since the verb *toru* literally means *to take* in English.

Semantic transparency, or how literal/idiomatic a MWU is, has often been utilized to identify collocations which have a higher learning burden and thus need extra study time. Clearly, the literal *eat breakfast* will be easier for a learner to learn in comparison to the idiomatic and difficult to decipher *fight tooth and nail*.

Yet another important criterion to consider is whether or not a collocation has balanced dispersion throughout English if a learner's goal is to master general English. If so, then that learner should only focus on collocations which occur in a balanced way among a variety of genres. Furthermore, if corpus data is utilized it needs to be confirmed that a collocation has stable dispersion over time, and is not dated, too modern, or occurring only during a specific time period in high numbers.

With Shin's approach, not only frequently co-occurring collocations are identified, but collocations with a higher learning burden than others (collocations which either have low L1 congruency or those which are semantically opaque) are as well. Or does the sheer number of collocations rule out any methodical approach to teaching, as Mackin (1978) claimed? If collocations are defined by frequent co-occurrence, how should we count such lexical co-occurrence? Should words be counted as *word types*, as Shin (2006) did, or would *word families* or *lemma* be more ideal? What exactly are the differences between these different ways of counting 'words'?

If one counts using word types, it would be counting of all words with distinct spellings separately with no regard for grammatical inflection. For instance, when the words *govern, governing,* and *government* are counted as 'word families', the three words would be counted together as one family connected by the stem *govern*. When counting as 'lemma', the verbs *govern* and *governing* would be counted as one and the noun *government* counted separately, and thus the count would be two lemma. However, when the words are counted as word types all three are counted separately with no attempt to consolidate data.

Counting using word families is quite different. A word family includes "a base word and all its derived and inflected forms" (Bauer and Nation, 1993, p. 11). For example, the word family for *govern* is represented by the headword *govern,* and represents *gov, governed, governing, government, governmental, governments, governor, governors, governorship,*

governorships, governs, govt, intergovernmental, misgoverned, misgoverning, misgoverns, and *ungovernable* (Heatley, Nation, & Coxhead, 2002).

In between word types and word families are lemma. A lemma, as defined by Nation and Meara (2002), is a "set of related words consisting of the stem and inflected forms that are all the same part of speech" (p. 36). For example, the verb run would be the lemma that represents the forms runs, running, and ran while the noun run would be listed as a separate entry.

But which of these ways of counting words is ideal for the research questions set forth in this thesis? Later in this study the rationale as to why lemma are the ideal way to count co-occurrence of words will be explained illustrated further.

Moreover, are positional and constituent variation (concgramming) truly important criteria to consider? If so, can currently available concordance software process data in a way that will help identify items most worthy of study? What would be an appropriate frequency cut-off for high-frequency collocations, and do the resulting identified items constitute a practical learning goal for direct study? These and a number of other important questions remain unanswered. The goal of this current research is to answer them, which will help traverse the barriers that prevent learners from obtaining collocational fluency.

Research Questions

1. What is a frequency data cut-off for lemmatized concgrams that results in a list consisting of 2-3,000 word families?

2. To what extent is corpus dispersion data useful for identifying MWUs that are deemed worthy of instruction by native English speaker intuition?

3. To what extent is corpus chronological data useful for identifying MWUs that are deemed worthy of instruction by native English speaker intuition?

4. To what extent is consideration for colligation an important criterion for identifying MWUs that are deemed worthy of instruction by native English speaker intuition?

5. What percentage of MWUs is deemed by experienced native speakers who teach English as a second language at university worthy of expanding beyond their most frequent exemplar to provide learners with useful information about how the items commonly occur formulaically?

6. To what extent is semantic transparency an important criterion to consider when attempting to identify collocations deemed worthy of direct instruction by native speakers?

7. To what extent is L1-L2 congruency an important criterion to consider when attempting to identify English MWUs deemed worthy of direct instruction by native speakers to Japanese learners?

8. To what extent is native speaker intuition useful in regards to high-frequency vocabulary usage in context creation?

9. Are there any correlations between Japanese university students' knowledge of MWUs most representative of high-frequency lemmatized concgrams and TOEFL score, item frequency or L1-L2 congruency?

Conclusion

This chapter highlighted how teachers and researchers agree on the importance of collocational fluency for second language learners. It also showed that, despite realizing this, collocations are still not focused on and that the lack of a comprehensive resource which accurately reflects natural language is to blame. This chapter noted a number of difficult questions that have yet to be answered, which pose as barriers towards solving this issue, and thus identified a clear research problem which needs to be solved. The research questions that aim to help solve this research problem were therefore identified and listed.

Chapter 2
Review of the Literature

Introduction

Collocational knowledge is a part of second language acquisition that learners must master, but what exactly is a collocation? In fact, collocations are quite difficult to define and identify. This literature review will discuss the various ways in which collocations have been defined and identified, and the virtues and limitations that each methodology entails. This chapter will also discuss what previous research says about the value of collocational knowledge, and the lack of collocational knowledge among learners throughout the globe. Previous research on the learning burden of collocations, the lack of research and resources, the large amounts of data which must be grappled with, criteria to use to identify collocations, and on the direct teaching of collocations will be discussed as well.

Scope of the literature review

This literature review will cover all pertinent areas of research necessary in regards to improving upon second language learners' collocational fluency. The main goal of this literature review is to highlight research which has defined and/or clarified the phenomena of collocation, pointed out the importance of collocational fluency, and identified a lack of knowledge of and/or resources which help develop collocational fluency in second language learners. It will also review research which specifies important criteria researchers should consider when attempting to identify useful collocations, and which items deserve direct teaching time.

This literature review will not examine native speaker acquisition of collocational knowledge because this study is focused on creating materials which would help ESL learners obtain collocational fluency, whom acquire collocational fluency in very different ways due to practical limitations of exposure. This review will also not examine studies which are concerned with how collocations are specifically stored in the brain. This current research has the practical pedagogical intention of identifying high-frequency collocations with the goal of teaching them directly to ESL students. It is a fact that the phenomena of collocation exists, native speakers possess such knowledge, and non-native speakers often do not. Simply identifying which ones occur frequently is the goal here, and how they are stored in the brain is beyond the scope of this study.

This literature review will also not delve into differences in collocations among the varieties of English for a number of reasons. First, since parts of this study will require native speaker judgments on language, accurate judgments cannot be made for a variety of English that is not one's mother tongue. Second, age and quality issues in regards to certain corpora have led to the decision to work with data from only one specific corpus, which is American English. Third, the target learners that this current research aims to help are Japanese, who are mostly taught the American variety of English due to the fact that the largest group of native English speakers in Japan are by far from the United States (see Table 1 below). Thus, it is only logical that this variety of English be examined.

Table 1

Foreign national residents by nationality in Japan (the top six native-English speaking countries) (Japanese Ministry of Justice, n.d.)

Country	Population
United States	49,979
United Kingdom	14,880
Canada	9,024
Australia	9,014
New Zealand	3,109
Ireland	1,039

Overview of vocabulary learning research

How does a person achieve fluency in a second language? Ellis (1985) stated that theories on second language acquisition abound, and that perhaps there are too many and that some may have been accepted as fact too soon (p. 248). Hadley (2001) discusses these theories by placing them on a continuum with empiricists on one side and rationalists on the other. Empiricist theories of language learning include Skinner (1957), who proposed that *Operant*

Conditioning is how learning occurs in humans. Positive and negative reinforcement by the community shapes the language that a learner will persist in using. From this theory's viewpoint, the human mind is a *tabula rasa* upon which pre-established accepted language patterns are imprinted upon. Rationalist theories, such Chomsky's (1957) *Universal Grammar,* rejected such empiricist theories, by rather insisting that humans are innately programmed to learn language.

In more recent years, other theories have been developed. Gasser's (1990) connectionist theory of language acquisition would fall close to Skinner's on the empiricist end of the continuum. It describes the storage of language in the brain as a network of interconnected units which are "strengthened or weakened in response to regularities in input patterns" (Gasser, 1990, p. 179). But again, on the other end would be Krashen's (1982) *Monitor Model,* which stated that conscious (*Learning*) acquisition of grammar rules does occur when a person learns a language, along with unconscious (*Acquisition*) learning as well.

However, despite the existence of contradictions between such theories, Larsen-Freeman and Long (1991) suggest that it would not be prudent to accept only one of these theories as omnipotent at this early stage in the field of language acquisition research. This is true today as well. For instance, despite the existence of a tremendous amount of research on language acquisition, there still is not agreement on a universal theory of language acquisition and many researchers still argue about the shortcomings of the theories mentioned above. Furthermore, new shortcomings are still being identified.

Now, considering that this current study's focus is the identification of the high-frequency collocations, which constitutes a major gap in the research, it is clear that Larsen-Freeman and Long (1991) were correct in suggesting that it would be imprudent at this stage to assume that any one theory should be accepted. It is thus clear that many questions still remain unanswered in this field, even some of the most important basic questions such as a general theory of language acquisition.

Within the study of second language acquisition, much research has been done on vocabulary acquisition. Researchers have shown that there is an order that words need to be learned by children (Anderson & Nagy, 1991), that a number of exposures to a word was necessary for a learner to truly master full knowledge of it and how it can be used (Nagy,

Herman, & Anderson, 1985), that while sometimes new vocabulary may not be being learning through incidental learning other types of vocabulary knowledge gains do occur (Waring & Takaki, 2003), and a wide variety of other discoveries in regards to this fundamental aspect of second language acquisition.

Such research had led to the development of a number of methods that are utilized to help second language learners acquire fluency. For example, the *Direct Method* has learners make discoveries about grammar inductively through adequate exposure to linguistics forms. This contrasts strongly with the G*rammar-Translation Method*, in which learners are explicitly taught grammar rules and must apply said rules and translate between the L2 and their L1. *Communicative Language Teaching* is another methodology which focused on the needs of learners to be able to accomplish their own communicative goals rather than focusing on the mastery of grammar with an emphasis on interaction, authentic language, and linking the classroom learning to real world experiences. Many more methodologies exist, each has their strong and weak points, and some are more useful for certain goals than others.

One particular methodology which is relevant to this current study is the *Lexical Approach*, a method pioneered by Lewis (1993), which focuses on the learner's ability to understand and produce lexical chunks of language. Lewis (1993) suggested that the linguistic phenomena of collocations may actually be the central organizer of language, and thus his theories are central to the questions that will be explored in this thesis. With the *Lexical Method*, words are presented to the learner in the form of the common chunks they usually occur as in instead of as isolated vocabulary. Certain words arbitrarily co-occur in these chunks which cannot be explained through logic, and thus more of a focus on mastering them instead of grammar or isolated vocabulary is called for since the learner will acquire the vocabulary and grammar indirectly via these chunks. Since its inception, computer technology has developed to the point where materials writers can use concordance software and corpora to identify such language, and have that inform what language they choose to focus on when creating materials for learners instead of simply relying upon their intuition and teaching experience. The way that such chunks can be accurately identified is through use of co-occurrence frequency data and is the ultimate goal of this current study.

However, so many questions regarding collocations still remain and it is clear that much more research is needed. Moreover, there is not even consensus on how to define a collocation, let alone identification of the common ones of the English. Shin (2006) and Cowan (1989) both stated that there is too much variability in researchers' definitions of 'collocation'. For instance, many researchers defined collocations by their tendency to frequently co-occur (Hoey, 1991; Jones and Sinclair, 1974; Firth, 1957). Others used syntactic structures (Gitsaki, 1996; Zhang, 1993). Some researchers even used a combination of both frequency data and syntactic patterning to identify collocations (Lesniewska & Witalisz, 2007).

This current study therefore aims to take a small step in first developing a methodology that will define and identify the common collocations of English and the formulaic language they most frequently occur in. Once accomplished, then such data could not only be used to inform pedagogy, but also to help add to data researchers can use to further develop overarching theories of language acquisition. This study will thus take an all-encompassing approach to defining collocation by frequency of co-occurrence. Literal collocations will be examined as well as idioms with the ultimate goal of identifying MWUs most representative of lemmatized concgrams. The justification for this all-encompassing definition will be revealed as this study progresses and its unique approach is explained. This study will also utilize a variety of other criteria to help pinpoint the exact items learners need to focus on the most.

The lack of collocational fluency

A lack of collocational fluency among second language learners seems to almost be a universal issue. "That learners have problems with collocations is a well-established fact" (Biskup, 1992; Bahns & Eldaw, 1993; Howath, 1996; Granger, 1998, Nesselhauf, 2005). Even as early as the 1970s, researchers wrote about the lack of this essential aspect of second language acquisition. Grucza and Jaruzelska (1978), Marton (1977), and Arabski (1979) all note that a large percentage of student errors are actually collocational errors.

Research shows that learners from a large variety of backgrounds struggle to obtain collocational knowledge. In Europe, second language learners struggle. Linnarud (1986) found that Swedish learners utilized collocations much less in comparison to native speakers. Biskup (1992) found that both Polish and German university students are lacking in collocational

fluency. Bahns and Eldaw (1993) found that approximately 50 percent of collocational phrases translations by German EFL students were incorrect. Nesselhauf (2003; 2005) also showed that German students have weak knowledge of collocations. Jaen (2007) showed that university students studying English linguistics from Spain had poor collocational knowledge as well. In the Middle East, second language learners find collocations difficult to learn as well. Fayez-Hussein (1990) found that Jordanian university students majoring in English could not provide the correct answer approximately 50 percent of the time when their collocational knowledge was tested. Keshavarz and Salimi (2007) found that Iranian EFL learners also have insufficient knowledge of English collocations. Asia learners have similar issues. Both Lin, Hsiao-Ching and Ho-Ping (2003) and Liu and Shaw (2001) found Taiwanese university students to also have limited collocational knowledge. Rogers (2013) found collocational fluency to be among the weakest scores of vocabulary depth knowledge that Japanese university students possess. Tseng's (2002) questionnaire even revealed that Taiwanese high school students actually knew little of even the concept of collocation.

We have known for some time now that collocational errors actually make up a very large percentage of second language learner errors in general (Arabski 1979; Grucza & Jaruzelska, 1978; Korosadowicz-Struzynska, 1980; Marton, 1977). Furthermore, this is not just a problem for lower level students. In fact, we have also known for some time now that even advanced level learners struggle with collocational knowledge (Brown, 1974; Channell, 1981; Cowie, 1978; Hausmann, 1984; Mackin, 1978; Rudzka, Channell, Putseys, & Ostyn, 1981). Unfortunately, the problem persists. In the 1990s, Bahns and Eldaw (1993), Biskup (1992), Gitsaki (1996) and Kjellmer (1990) all noted that high level learners have limited collocational fluency. More recently, Wang (2001) found collocational knowledge to not increase relative to academic levels. Liu and Shaw (2001) and Nesselhauf (2005) found this as well, in that the number of years learners studied English was shown to have no effect on their collocational fluency in her study.

But what makes collocational fluency so difficult to acquire? This question will be answered in the following section.

On the learning burden of collocations

Obtaining collocational fluency is not an easy task to accomplish. Researchers have been cognizant of the difficulty of mastering collocational fluency for some time now. The same is true for idioms/formulaic sequences. Wilkins (1972) stated that "the appropriateness of idiom to situation is very difficult to master" (p.128). Similar opinions continue to this day. The term 'collocation' itself is limiting, but this current study uses it to refer to not only the collocates themselves, but also the formulaic sequences they commonly occur in, whether they be literal, figurative, or idiomatic. As this study progresses, the rationale behind why this is necessary to discuss and consider all of these kinds of 'collocates' together will become more and more salient. In fact, this current study not only aims to identify useful collocations, but goes further in also identifying the common formulaic sequences they commonly occur in (the MWUs most representative of 'lemmatized concgrams', or 'collocates'). The goal is not to define or redefine 'collocations', but rather to identify what items learners need to study to help them master collocational fluency in the most efficient way.

Bahns and Eldaw (1993) found that German students' productive knowledge of collocations in particular was limited. Jaen (2007) also found productive knowledge to be significantly less than receptive knowledge in students from Spain. Nesselhauf (2005) found that when writing under time pressure, second language learners do not use collocations to the same extent that natives do. These issues are easy to understand because productive knowledge will always lag behind receptive knowledge. However, in regards to the other specific aspects of collocational knowledge that learners struggle with, there are actually a variety of issues that serve as barriers to students mastering this knowledge.

Bahns and Eldaw (1993) specified that collocational knowledge significantly lags behind general vocabulary knowledge. But why? By far, the sheer number of collocations makes it probably the most challenging aspect of mastering vocabulary depth. While it is difficult to pinpoint exactly how many collocations a native speaker has in their lexicon, some researchers have estimated the number to be in the hundreds of thousands (Hill, 2000). For instance, Davies' (2010) collocation list had 50 collocations with the lemma *water* having more than 500 occurrences per 425 million tokens, and while the value of the higher frequency items, such as *drink/water* (3,099 occurrences), is clear, even items with much lower frequencies, such as *splash/*water (592 occurrences) have obvious value. Gitsaki (1996) agreed, stating that "one of

the main reasons the learner finds listening or reading difficult is not because of the density of new words, but the density of unrecognized collocations" (p.54).

Hill, Lewis, and Lewis (1996) highlight an additional issue that makes obtaining collocational fluency a difficult task: its complexity. They wrote:

> "collocation is never as simple as it seems - sometimes the adverb must come in front of the verb, sometimes it must come after, and sometimes either position is possible with very similar meanings. Some adjective + noun or verb + noun combinations are much more common if they are used in the negative; perhaps some of the verbs are used with the headword mostly when it is literal, others mostly when it is more metaphorical. Very rarely are the lines between two 'different' uses of this kind clear. (p. 116)

Learners struggle with this complexity, and will often make errors by overgeneralizing, or substituting a generic term for something that is normally represented by a more arbitrarily fixed term. Farghal and Obiedat (1995) tested Jordanian learners on their collocational knowledge, and found such issues, and gave examples such as learners producing *heavy tea* instead of *strong tea*. Fayez-Hussein (1990) found that such errors accounted for 38.3 percent of the collocational errors made in his study.

Another issue that Moon (1997) noted is how the non-compositional nature of how collocations are formed necessitates that learners recognize, learn, decode, and encode them as holistic units, and this significantly adds to their learning burden. The results of Jaen's (2007) study on collocational knowledge of university students in Spain also showed that the arbitrary nature of how collocations are formed to be problematic and responsible for the students' difficulties with them. Laufer (1990) also made a point to mention this issue, referring to it as the "rulelessness of collocations" (p. 147). Fayez-Hussien (1990) gives the example of *several thanks* vs. *many thanks*, and the inability for learners to use any kind of logic to determine why one is appropriate and the other is not.

Furthermore, how semantically bonded a collocation is, or in other words how high of a chance one word has of occurring with another, seems to also affect its learning burden. Nesselhauf (2003) found that the highest rate of errors occurred with collocations that had a

medium degree of restriction, or more specifically, when the verb cannot "be used with every noun that would be syntactically and semantically possible" (p. 233), or the situation where the verb can only combine with a limited number of nouns. She gives examples such as *exert influence,* and how there is a medium degree of restriction in that other noun possibilities exist (*exert control*) but not others (*exert rights*). Keshavarz and Salimi (2007) pinpointed that Iranian students particularly struggled with restricted collocations. Howarth (1998), Huang (2001), and Nesselhauf (2003) noted a similar weak point in that in their studies, learners tended to make errors with restricted or semi-restricted collocations. Biskup (1992) found the same issue. In his study, Polish students only produced acceptable restricted collocations correctly 22.6 percent of the time while German students only produced them correct 16.6 percent of the time. Liu and Shaw (2001) found this issue as well. In their study, learners produced significantly less free combinations in comparison to 'pre-fabs'.

Learners seem to also struggle with particular types of collocations. Hsu and Chiu (2008) found that the learners in their study never produced adverb/adjective collocations, and recommended that teachers focus on such items. Liu (1999) found verb-noun errors to be the most numerous type of collocational error that Chinese college students made. Liu (2002) found that 87 percent of errors Taiwanese students made were verb-noun combinations, and in 93 percent of them the verb was the problem. Moon (1997) found phrasal verbs to be problematic for learners. Nesselhauf (2005) discovered that specific semantic groups of verbs were difficult for her students. Her German students particularly struggled with the verbs *achieve, reach, acquire, obtain,* and *gain.* In general, many researchers also cite how semantically transparent a collocation is, or in other words how literal/figurative a collocation is, can have an effect on its learning burden. Gitsaki (1996) cites semantically opaque examples, such as 'foot the bill' and 'high explosive', and their obvious potential to mislead. Thus whether an item is semantically transparent, and whether students are aware of this, can affect a collocation's learning burden.

There are a variety of consequences of weak collocational knowledge. Durrant and Schmitt (2009) highlighted a variety of research which shows that learners have a tendency to overuse certain phrases, especially if they are frequent, neutral, or exist as a cognate in their L1. Gitsaki (1996) highlights how "In English people 'draw conclusions' while the Greeks 'bga;zounsumpera;smata' [take out conclusions]" (p. 3-4). Fayez-Hussein (1990) gives the

example of how students produced *pipe water* instead of *tap water*. But why do learners make such errors?

As mentioned above, L1-L2 congruency, or or how similar/dissimilar a collocation or MWU's translation is in the learner's native tongue, is a major factor influencing the learning burden of collocations. Both Nesselhauf (2005) and Fayez-Hussein (1990) found that approximately 50% of collocational errors were due to L1 influence, and thus such items should receive more teaching time. Chan and Liou (2005) found that 38 percent of collocational errors were due to L1 influence. Chen (2002) found L1 interference to be a common source of errors by Taiwanese high school students. Al-Zahrani (1998), Bahns (1993), and Biskup (1992) all call for increased emphasis on non-congruent collocations.

Mutual information is the likelihood that one word will occur with another with consideration for word frequency. For instance, *crux/matter* have a high M.I. score of 6.15. Durrant and Schmitt's (2009) study also showed that learners significantly underuse collocations with a mutual information score of over seven (see section 2.9.2 for more information about the use of mutual information data). They also found that learners do not use as many low-frequency collocations as natives. Patterns of underuse and overuse of certain collocations has also been noted by DeCock, Granger, Leech, and McEnery (1998), Granger (1998), Lesniewska and Witalisz (2007), and Lorenz (1998).

All of these studies highlight how the obtainment of collocational fluency is not being achieved. Even advanced level learners still lack this essential skill. The problem is not simple by any means, either. The above studies also highlight how particular types of collocations have a higher learning burden than others, such as arbitrarily bonded collocations, restricted collocations, L1-L2 incongruent collocations, low-frequency collocations, among others.

However, although the studies in this and the previous section make it is quite clear that second language learners across the globe struggle to obtain collocational fluency in English, and in particular have difficulty with certain types of collocations, previous research has yet to comprehensively pinpoint the extent to which certain aspects of collocational fluency cause difficulty for learners. This current study's research question 9 aims to fill this gap in the research by judging in fine detail Japanese university students' general collocational knowledge,

and whether certain aspects of collocations, such as frequency and L1-L2 congruency, play a factor in increasing their learning burden.

On approaching and defining collocations

What is a collocation?

As mentioned earlier, a variety of terms and criteria have been used by a multitude of researchers to define and operationalize the term 'collocation' and how collocations formulate into MWUs. In fact, it is quite difficult to distinguish 'collocations' from phrasal verbs, prefabricated patterns, idioms, and so on. The approach this study will take to define 'collocation' is simply words that frequently co-occur, but in actuality, this study will also discuss collocations beyond the concept of two words frequently co-occurring adjacently, but also as the MWUs they commonly co-occur in. This is achieved by identifying collocations and the MWUs they most commonly occur in as *lemmatized concgrams*[3]. For example, when corpus data points to *take/break* co-occurring frequently, the lemma families of *take* and *break* are all examined for the various ways in which the two lemma co-occur in various different ways with other words to give a co-occurrence count which better reflects natural language (*take breaks, took a break, taking breaks,* etc. are all counted as co-occurrences of the lemma *take* and *break*), and finally the MWU most representative of how these two lemma co-occur (*take a break*) is identified using the concgramming methodology and specialized concordance software.

To understand the strong and weak points of all of the valid methods of defining collocations, the three main approaches to researching collocations must first be discussed. The three main approaches to studying collocations are semantic, structural and lexical. In the *semantic approach*, collocations are defined as being predictable by their semantic features (Robins, 1967). This approach aims to explain why particular lexical items occurred only with certain others. Gitsaki (1996) pointed out that a weakness of this approach is that, "There is a large number of idiosyncratic co-occurrences or combinations that are arbitrarily restricted…they are left unexplained and marginal by semanticists" (p. 35). Gitsaki (1996) listed some examples,

[3] A more detailed explanation of the term *concgramming*, this methodology and rationale why such an approach is being taken will be given in subsequent sections of this thesis.

such as how *kick the bucket* and *blond hair* can only be used when referring to humans (p. 33). Lewis (2000) agrees, in that trying to use semantics to explain why certain words co-occur leads to, at best, "half-truths" (p. 13).

Meanwhile the *structural approach* utilizes grammatical patterns to explain collocation, and proponents believe that collocation is influenced by structure. Mitchell (1971) proposes that collocation be studied within these "grammatical matrices" (p. 48). Gitsaki (1996) agrees, in that his study of 275 Greek learners of English at three separate proficiency levels showed that the learners did not once use a number of particular collocation patterns, such as *adverb+adjective,* and that these were avoided due to their structural and syntactic complexity and relative infrequency in English. However, Hill (2000) distances himself from "previously cherished structuralist ideas" (p. 48) and believes that instead of breaking down language into smaller and smaller categories, we should try to view language in the largest units possible. Thus, this statement leads us to the lexical approach.

Regarding the *lexical approach,* Halliday and Sinclair (1966) begin to consider lexis as separate, but complementary to grammatical theory. They believe that it is necessary to consider collocation's influence on the organization of language because grammar alone was not enough to determine which lexical item would occur due to the idiosyncratic nature of collocations. Halliday (1966) cites how word choice can also be specified by collocational restrictions, in addition to structural and semantic limitations (p. 152). He gives the example of how *strong* is a member of a lexical set with *tea*, and *powerful* is a member of a lexical set with *car,* which cannot be explained by the structural or semantic approaches. Lewis (1993) stated that language "consists of grammaticalised lexis, not lexicalized grammar" (p. vi). The *lexical approach* thus views lexis, and not grammar, as the overarching engine that organizes language.

Each of these approaches has its strengths and weaknesses, and their usage depends on the type of research being conducted. However the lexical approach does have advantages over the semantic and structural approaches, as is evident in Table 2 below. The verb *play* in fact has many different meanings in English, and the examples below highlight them. The most typical usage people will think of is its usage to describe participation in a game or sport or the use of a musical instrument. However, one can also *play politics* or *play a character* in a film. *Politics/a character* and *sports/musical instrument* are both nouns, so the structural approach would

identify the pattern *play + [noun]*. However, one can also *play him/herself* (the slang usage that means to make a fool of oneself), and thus the structural approach would miss all instances of *play + [pronoun]*.

The semantic approach also fails to cover all usage of the verb *play*. Through the semantic approach, logic is used to understand the verb's usage. You *play* something that you need to practice, such as a musical instrument, a sport, or even a character. However, the logic of this approach fails with *play politics* and *play him/herself.*

However, the lexical approach can identify all of the above mentioned co-occurring patterns by only focusing on frequency of co-occurrence. Although each approach has its place in collocation research, the above examples highlight the significant advantages of the lexical approach for the goal of this particular study.

Table 2

The approaches' ability to identify common collocates of the verb 'play'

Collocates of the verb *play*	Semantic Approach	Structural Approach	Lexical Approach
play {sports}/{instruments}/ {music}/{games}	O	O	O
play politics / play a character	X	O	O
play himself / play herself	X	X	O

Thus, a lexical approach will be taken to ensure that all important collocates are identified. This study will therefore define collocations in the traditional sense as words that have a high frequency of co-occurrence (Biber et al., 1999; Shin, 2006). This study will also include some aspects of a structural approach, and the justification for this will be discussed later in this study. However, it is still unclear how the criterion of frequency should be applied. For

example, what frequency cut-off should be utilized to identify only high-frequency collocations worthy of direct study? Research question 1 in this current study addresses this issue.

Types vs. lemma vs. word families

When we define collocations by their tendency to frequently co-occur, word combinations such as *jury's verdict* are identified. However, combinations such as *of the* are also identified. Should such grammatical combinations also be considered as 'collocations'? Does teaching *of the* have value to a second language learner? Shin (2006) believed that it does not, explaining that an important criterion of collocation identification is that it needs to be a meaningful unit, or in other words, grammatically well-formed. One way of accomplishing this is by only considering content words as collocations (nouns, verbs, adjectives, and adverbs), as did Woolard (2000). Ackermann and Chen (2013) also limited their dataset in a similar way by only examining verb-noun, adjective-noun, adverb-adjective, and adverb-verb formulations. This study will thus only consider *pivot words* and collocates which are either a noun, verb, adjective, or adverb. *Pivot word* refers to the focal word in a collocation (also called a 'node'). In a 'collocation', it is accompanied by its 'collocate', or the word/words accompanying the pivot word (Shin, 2007). For example, a search for collocates for the pivot word *eat* may bring up words such as *lunch,* while a search for collocates for the pivot word *lunch* may bring up words such as *break.* Together, a pivot word and a collocate make up a 'collocation'.

Biber et al. (1999) deemed collocations to be two-word phrases which co-occur, distinguishing them from idioms and lexical bundles. With second language learners in mind, this actually is not ideal. Take the collocates *crux/matter* for instance. These two words clearly collocate, but never simply as a two-word phrase. They always collocate within the larger chunk *crux of the matter.* Therefore, limiting the definition of 'collocation' to two-word phrases excludes items which clearly collocate. Researchers such as Conzett (2000) improved upon the definition of collocations by considering two or more frequently co-occurring words as 'collocations', or what is more commonly referred to as MWUs.

Defining MWUs is actually problematic as well. A variety of terms have also been used to describe them, such as 'combinations of lexical items' (Korosadowicz-Struzynska 1980), 'conventionalized language forms' (Yorio, 1980), 'prefabricated language chunks and routinized

formulas' (Nattinger & DeCarrico, 1992), 'phrase patterns and sentence patterns' (Twaddell 1973), 'word associations' (Murphy, 1983), 'fixed expressions' (Alexander 1984; Kennedy, 1990), and 'formulaic language' (Wray, 2002). To categorize different types of collocation is also problematic in that they often overlap in what they describe.

Furthermore, there is the issue of how words should be counted. Should they be counted as word types, as Shin (2006) did? With his word counting approach, all words with distinct spellings were counted separately with no attempt to consolidate data. This is in contrast with counting words as *lemma* or *word families*. Such a method would be successful in accurately identifying *crux of the matter*. However, in certain circumstances this is not the most ideal way to 'count' frequency of co-occurrence. The reason why is the amount of collocations that exist in a language can be in the hundreds of thousands (Hill, 2000; Pawley & Snyder, 1983) and there is a clear need to consolidate data in some way if the goal is to identify collocations worthy of direct study. The focus of this study is to identify collocations to directly teach to second language learners, and thus efforts need to be made to grapple with the copious amount of collocations that exist. Realistically speaking, there simply is not enough classroom time to teach every collocation. Fortunately, options are available to help consolidate data, such as by counting words as *word families* and also as *lemma.*

However, there are issues counting using word families. Webb and Nation (2008) remark that if learners demonstrate knowledge of a headword, there is an assumption that they also have receptive knowledge of the rest of the word family. However, depending on the goal of the study, using word families may not be ideal. For instance, Schmitt and Meara (1997) actually found that Japanese high school and university students had poor English affix knowledge. Daulton (2008) agreed, stating that it is "imprudent to assume that Japanese learners can extend word knowledge within word families" (p. 120).

Furthermore, teachers have the practical goal of teaching high-frequency vocabulary. Ideally, such vocabulary should be taught along with its high-frequency collocations in the form of MWUs. When a teacher selects a word worthy of teaching using word families, one or more examples must be given. Let us imagine a situation where a teacher needs to teach a word within the word family for *govern.*

By using frequency data from the COCA and counting frequency as lemma, Table 3 below shows that the lemma in the word family which have the highest frequencies are *government, governor, govern,* and *governmental.* When relying on native speaker intuition, these four words are of clear value to be taught directly to second language learners while other words in the word family are considered to have either marginal or low value for direct teaching. Table 3 below also lists those words most frequent collocations. The corpus data reveals the collocation *government/federal* as the most frequent. However, very different but still valuable collocations also occur with the other lemma in the word family with vastly different frequency counts. So, what should teachers do when presented with the task of teaching one of the words in this family? Should they rely simply on the most frequent collocates in this family? If so, and they provided five examples for the learner, they would all be collocates for *government,* and all other collocates for the other common lemma in the word family would be excluded. This data set clearly shows how word families have the potential to be overly inclusive.

Table 3

High-frequency collocations for the four most frequent words in the word family for 'govern' according to the COCA[4] *(top frequencies in bold)*

Pivot Word (frequencies)	Collocates (frequencies)
government (220,945)	**federal (15,542)** local (5,022) officials (4,336) agencies (2,721)
governor (42,639)	former (2,438) Republican (1,322) lieutenant (890) Democratic (750) office (748)
govern (8,552)	rules (721) laws (424) law (250) country (240) regulations (238)
governmental (4,374)	agencies (232) affairs (186) action (115) committee (113) institutions (112)

[4] Excluding proper nouns such as person's names and states

This data makes it salient that a more practical alternative to types and word families would be to count words using lemma, which is the procedure that this study has adopted. For example, consider the lemma pair *take/walk.* First of all, should all of the MWUs of the lemma pair *take/walk* really be counted separately as types? Aren't the MWUs *take a walk, take walks, took a walk,* and *taking walks* essentially part of the same group? Practically speaking, considering the copious amount of data such as that which this study must deal with, the answer is yes. In addition to the advantage of having less items to study, counting co-occurrence in such a way leads to frequency counts that better resemble natural language. For example, in the COCA (Davies, 2008) the types *take/walk* have co-occurrence frequency of 1,125, while the lemma pair *take/walk* have nearly twice that at 2,049.

Moreover, imagine a learner studies the MWU *took a walk,* and then later *take a break.* Would that learner be able to comprehend *take a walk* without directly learning it? The answer is there is a high probability that they would not have to because the affix knowledge necessary to comprehend the inflections that noun, verb, adjective, and adverb lemma comprise pose a very low learning burden. It is even more clear when noun lemma are considered. For example, learners clearly do not need to learn *powerful engine* and *powerful engines* at different times. All they simply have to do is master the general rule of how to pluralize nouns in English to have sufficient enough knowledge to comprehend such items.

On semantic transparency: Are literals, ONCEs, figuratives, core idioms all 'collocations'?

Semantic transparency, or how literal/figurative a collocation/MWU is, has often been utilized to identify collocations. Van der Meer (1998) discussed the distinction between what some refer to as *free combinations,* or formulations which are not preconstructed but are semantically literal, being different from the distinct category of collocations. Researchers such as Moon (1994; 1997) believe such literal formulations are not worthy of direct study by second language learners. However, rather than limiting oneself to rigid definitions of the term *collocation,* this study aims to focus on what word formulations are of value to teach second language learners. With such a goal in mind, not including free combinations as collocations becomes problematic because issues such as L1-L2 congruency come into play. Fayez-Hussein

(1990) actually found that 50 percent of collocational errors was due to L1 interference, and thus a literal collocation can still pose a high learning burden.

It is true that collocations can be categorized in a range of learning burdens using semantic transparency, such as how Grant and Bauer's (2004) taxonomy breaks formulations down into literals, collocations with one non-compositional element (ONCEs[5]), figuratives, and core idioms. With everything relative, semantically opaque collocations such as ONCEs, figuratives, and core idioms do pose a higher learning burden than literals. However, other researchers insist that literal collocations also be taught directly in addition to the issue of L1-L2 congruency. Nesselhauf (2005) found that students sometimes assign literal meaning to collocations with a figurative meaning, and vice-versa. Gitsaki (1996) noted that such collocations even show "a certain degree of syntactic frozenness and resistance to lexical substitution" (p. 49). Thus, this thesis defines collocations without excluding literal formulations since they do have the potential to be of value to be taught to learners.

However, where high-frequency MWUs fall on a spectrum of semantic transparency has yet to be determined comprehensively in previous literature, and thus this will be addressed by research question 6 in the current study.

On concgramming, MWU length and colligation

As discussed earlier, it is clear that it is not ideal to simply count words as types and provide learners with such word sequences to study. Such a method does not result in counting co-occurrence in a way that reflects natural language. Counting the occurrences of collocations does present itself with some issues, such as constituent variation. For instance, researchers such as Renouf and Sinclair (1991) used syntactic frameworks to grapple with discontinuous sequences. Wilks (2005) used a more advanced approach by utilizing *skipgram* searches, which can handle constituency variation. For example, it could be argued that *close friends* and *close childhood friends* should be counted together due to the fact that it is essentially the same collocation albeit with an adjective added. Cheng, Greaves, and Warren's (2006) *concgramming*

[5] Certain formulae are partially non-compositional, such as how in the collocational phrase *short and sweet*, *sweet* represents the one non-compositional element.

method was also a major step forward, in that it counted co-occurrence not only with consideration for constituent variation, but also positional variation. They stated that "searches which focus on contiguous collocations present an incomplete picture of the word associations that exist" (p. 431) in that the majority of the collocations they found in their study were non-contiguous, showing both constituency and positional variation, as is evident in Table 4 below.

A *concgram*, as defined by Cheng, Greaves, and Warren (2006), "constitutes all the permutations of constituency and positional variation generated by the association of two or more words" (p. 411). *Constituency variation* (AB, ACB) involves a pair of words not only co-occurring adjacent to one another (*lose weight*) but also with a constituent (*lose some weight*). *Positional variation* (AB, BA) refers to counting total occurrences of two or more particular lexical items that includes occurrences on either side of each other. Thus *provide you support* and *support you provide* would both be included in the total counts for a MWU concordance search for the lemma *provide* and *support*. Table 4 below shows the first four results of an actual concgram search for the lemma *provide* and *support*. This data is sourced from the COCA's online interface, which allows for lemma concgram searches and provides snippets of the sentences these concgrams are occurring in.

Table 4

A sample of data from the COCA for a concgram search for the lemma 'provide' and 'support'

...low-cost measures, the United States can extend the same lifesaving **support** that it has **provided** to the little boy in a rural, dusty village to the working-age woman living...
...it, then provide technical support to assist them. This **support** can usually be **provided** through a single phone call or demonstration. If needed, seek assistance from school...
...losing those aid dollars that we need in order to get **support** when Pakistan does **provide** it, which is real and does help us in the case of drones to...

...for low-income adults in occupational programs as well as financial **support** to colleges to **provide** support services for such students. States and colleges interested in adopting a model similar...

However, simply identifying lemma pairs that co-occur frequently is insufficient to provide learners with specific items to study. For instance, *take/walk* collocate, but it is not enough to simply expose students to this lemma pair. Rather, a more specific example of how the two collocate as a MWU needs to be identified. Is it *taking walks, took walks, take a walk*, etc.? Thus steps are required to identify the MWU most representative of that lemmatized concgram. This is accomplished via concordance software, such as *AntConc* (Anthony, 2011). With such software, concordance data from a corpus can be processed to identify the MWU most representative for a lemma pair. When 500 example sentences containing both the lemma *provide* and *support* from the COCA were processed with *AntConc*, it is revealed that *provide support* is the most common MWU that occurs. Table 5 below shows the top three MWUs for this lemma pair.

Table 5

Top three MWUs for the lemma provide and support found after examining 500 concordance strings in the COCA

MWU	Frequency
provide support	55
support provided	39
support provided by	32

Concgramming has significant advantages when the goal is to identify MWUs most representative of high-frequency collocations. Attempts to identify MWUs that are not done as

concgram searches thus have the potential to produce results which do not accurately reflect natural language. Unfortunately, much of the previous research that aimed to identify high-frequency MWUs was actually conducted without consideration for positional or constituent variation (Biber, Conrad, & Cortes, 2004; Shin, 2006; Simpson & Mendis, 2003). Therefore, there is a clear gap in the research that this study aims to fill.

Furthermore, another pertinent question is whether a MWU identified as most representative of a lemmatized concgram should go beyond the pivot and collocate. For instance, should an identification method stop at *take a walk* or should it extend beyond this to identify *take a walk to*? An example of the need for such an approach can be seen in Table 6 below. The lemma *come* and *term* co-occur, but as evident in the table below, only *come to terms* is identified as the most frequently occurring phrase in which the two lemma co-occur in a corpus data search for these two lemma. However, *come to terms with* occurs nearly as much as *come to terms* but is not identified by corpus data (see Table 6) as the most frequently occurring because it occasionally occurs in less common formulations. Sometimes a sentence ends in *come to terms*, sometimes *come to terms* is followed by an interjection, and sometimes the rarer *come to terms on* occurs instead of *come to terms with.* Examples of such instances can be seen in the raw concordance data from the COCA utilized in this study below (Davies, 2008).

...others are ready to settle disputes and come to terms.

And hoping you and Peter might come to terms - that is –...

...they will come to terms on an Israeli-Palestinian accord.

Such extending of MWUs beyond their pivot and collocate needed in this current study is not actually possible with available concordance software. Thus, a native speaker must be relied upon to extend the sequence beyond the most frequent MWU to its left or right when the native speaker judged any additions to be part of the natural unit. This is possible by having native speakers rely on their intuition to only add strings to the core formulaic sequence that truly represented common usage, but that also provided learners with useful information. While somewhat subjective, practically speaking such a method does improve upon the ability to provide learners with useful information on how collocations are typically used. In fact, Simpson-Vlach and Ellis (2010) found experienced native speaker intuition to be an essential, valid and reliable criterion in selecting useful formulae in their study.

Table 6

MWUs identified from 500 example sentences in which the lemma pair 'come' and 'term' both occur in

MWU	Occurrences in 500 sentences
come to terms	243
come to terms with	229
to come to terms	133
to come to terms with	129
coming to terms	96
coming to terms with the	86
to come to terms with the	44
come to terms with [pre-nominal possessive pronoun]	28
coming to terms with the	26

There is also the question of how long exactly should a MWU be? In this thesis, the maximum length of a MWU is set at seven words long. The rationale for this length stems from findings on typical human memory limitations (Miller, 1956). In reality, this parameter was overkill in that the vast majority of MWUs identified were two to four words long, but it was utilized to ensure that no data was excluded.

Colligation, or the counting various lexical items that can easily substitute for one another as grammatical categories (Gitsaki, 1996; Renouf & Sinclair, 1991), is another important criterion for MWU identification which there is a lack of research. As discussed earlier, this would fall into the structural approach to understanding collocations. An example of colligation is counting the collocates *early* and *century* as *early* [*year*] *century* when they occur with years, which would account for instances, such as *early twentieth century*, *early nineteenth century*, etc., together. Table 7 below shows the advantage of processing corpus data with consideration

for colligation. One thousand example sentences were collected from the COCA (Davies, 2008), and a concordance search identified the MWU most representative of how *century* and *earlier* occur together. One search was done with consideration for colligation, replacing every instance of a year with the marker *[year]*. By considering colligation, the top MWU identified was shown to have nearly double the frequency in comparison with the top MWU identified without consideration for colligation.

Table 7

A comparison between two MWU searches, one with and one without consideration for a specific type of colligation

Without consideration for colligation		With consideration for colligation	
% of occurrences in 1,000 example sentences	MWU with co-occurrence of *century* and *early*	% of occurrences in 1,000 example sentences	MWU with co-occurrence of *century* and *early*
10.7%	*century earlier*	19.2%	*early in the* [*year*] *century*
9.5%	*a century earlier*	10.7%	*century earlier*
8.5%	*early in this century*	9.7%	*early* [*year*] *century*
7.3%	*early in the century*	9.5%	*a century earlier*
6.4%	*centuries earlier*	8.5%	*early in this century*
5.0%	*early in the 20th century*	8.3%	*early as the* [*year*] *century*
		8.3%	*as early as the* [*year*] *century*
		7.3%	*early in the century*
		6.4%	*centuries earlier*

However, depending on the goal of the research, colligation also has the potential to create more problems than it solves. For instance, when major content word categories, such as nouns or verbs, are replaced with colligational markers, the limitations of how a MWU can be formulated may not be conveyed to the learner. Take the colligational framework *[adjective] tea* for instance. Typical examples such as *hot tea*, *brown tea* or *strong tea* are perfectly logical, but it becomes very difficult to explain why *powerful tea* is not an option. Due to this idiosyncratic way collocations sometimes occur, grammar alone is not sufficient to determine which lexical items co-occur (Lewis, 2000). This is why it is better to rely mainly on the lexical approach. However, colligation can still be a useful criterion to consider when attempting to identify high-frequency MWUs. Yet how this criterion can be implemented and the extent of its value remains to be seen. Thus this thesis aims to clarify the value of specific types of colligational searches and will provide examples of the type of data that results from such consideration in research question 3.

The importance of collocations

Nation (2001a) stated that a variety of knowledge is necessary to truly 'know' a word. This 'vocabulary depth' knowledge includes not only includes semantics, pronunciation, orthography, word parts, concepts, associations, grammar, constraints on use, but also a word's possible collocates. A number of researchers believe that collocational knowledge is of significant value for the language learner. In fact, we have known about the value of collocational fluency for some time. Bolinger (1968) argues that we learn and memorise words in chunks. Later, he also argued that most of our "manipulative grasp of words is by way of collocations" (Bolinger 1976, p. 8)". Twaddell (1973) stated that teaching phrase-patterns and sentence patterns from the early stages of L2 learning may help vocabulary expansion. Among the other early advocates for the importance of collocations in L2 learning and their inclusion in L2 teaching is Brown (1974).

Collocational fluency has been referred to as a "decisive factor in developing fluency" (Almela & Sanchez, 2007, p. 37) and awareness of it a matter of "first-rate importance" (McCarthy, 1984, p. 21). Durrant and Schmitt (2009) state that "competent use of formulaic sequences is an important part of fluent and natural language use" (p. 157). Collocational

fluency is not just important for advanced-level language processing, either. Kjellmer (1987) stated that "collocations are indispensable and ubiquitous elements of any English text" (p. 133). Saville-Troike (1984) believes they are essential even in early stages of language learning.

In the past, fluency in formulaic language was considered to be of marginal importance (Ellis et al., 2008). However, in recent years a number of researchers have changed their view of its importance considerably. Some researchers go even further, asserting that collocations function as a central mechanism of how language organizes itself (Hoey, 2005). Lewis (1993) refers to this concept as *grammaticalised lexis.* However, even if a stance is taken that collocation plays a more minor role than that, many researchers still feel that mastery of its knowledge is essential for a learner to be considered fully fluent in a langauge (Bahns & Eldaw, 1993). Cowie (1992, p.10) agreed, stating that "it is impossible to perform at a level acceptable to native users, in writing or speech, without controlling an appropriate range of MWUs". Ellis (1997) gives the example of how the sentence *I wish to be wedded to you* is syntactically possible, but clearly unnatural from a native speaker's perspective.

Learning collocation in comparison with isolated words has been found to actually be easier (Ellis, 2001; Lewis, 2000; Taylor, 1983). For example, Bogaards (2001) found that multiword expressions containing familiar words were retained 10% more than completely new single words immediately after a learning session and also 12.1% more in a delayed posttest 3 weeks later. But why are they easier? Laufer (1988) stated that collocations are useful in a variety of levels of vocabulary acquisition and self-learning strategies. Schmitt (1997) explains how this is possible by presenting a number of different mnemonic strategies that can be used by learners. For example, a word in a MWU can serve as a mnemonic hook to help the learner remember the meaning of other words in the MWU that they have forgotten.

Take the word *spine* for example. Let us imagine a situation where a Japanese student learns the translation of the isolated word *spine.* Then, let's imagine another student who learns the MWU *injure my spine.* Now, let's imagine both students encounter the word in a reading passage without the word *injure*, and that both students have forgotten the translation. However, imagine that the student who studied the MWU remembers that *spine* occurs as *injure my spine* despite still not remembering the meaning of *spine.* Now, let's say this student understands the meaning of *injure.* From the meaning of this word, they can imagine that *spine* must mean some

part of a person's body. Through this mnemonic hook, their brain will be able to make the jump to remember the Japanese translation of *spine,* while the student who studied the word in an isolated matter is left with no alternative but to give up. Furthermore, each time that the student who learned the MWU makes that jump in their brain on their own, the connection between *spine* and its translation becomes stronger and stronger, and eventually they will not need to rely on *injure* as a hook. In this way, parts of a MWU has the potential to aid learners in memory retrieval, thus strengthening connections and making learning more efficient overall.

The lack of collocation research and resources

Collocations are also quite difficult to acquire because there is a lack of focus on directly teaching them. This stems from a lack of comprehensive resources. Nesselhauf (2005) wrote that "suggestions as to which individual collocations or groups of collocations that should be taught are scarce" (p. 254). The reason why there is a lack of resources is there is a lack of comprehensive research. She also noted that much of the previous research does not go beyond simply stating that more emphasis on teaching collocations is needed. Thus, one reason why practitioners do not emphasize collocations despite being aware of their importance is that there are still very few studies that identify which are the most frequent (Durrant & Schmitt, 2009), and/or the studies that have been conducted all lack in comprehensiveness or are flawed in some way. This has resulted in the direct teaching of collocation being "marginalized in the language curriculum" (Wood, 2004, p. 28).

One of the problems is the fact that much of the previous research has limited its scope to a specific type of collocation or MWU. For instance, Biber, Conrad, and Cortes (2004) only found 172 'lexical bundles', limiting themselves by a very conservative cut-off of 40 occurrences per million tokens and only considering four-word sequences. Simpson and Mendis' (2003) search for fixed, institutionalized, semantically opaque, academic idioms only identified 238 such items. Aghbar (1990) and Bahns and Eldaw (1993) only examined verb-noun collocations, while Channell (1981) only examined adjective-noun collocations. These studies produce results in stark contrast with claims that there are hundreds of thousands of collocations in a native speaker's lexicon.

While there is an abundance of collocation dictionaries available, they tend to present users with too much information. For instance, Kjellmer's (1994) collocation dictionary contains over 85,000 entries, and pinpointing the most useful collocations from such a large dataset is clearly not an easy task. This lack of resources that specify useful collocations is thus clearly connected to the sheer number of items researchers must deal with. Shin's (2006) study was a good first step in alleviating these issues, but his study was limited by only examining the most frequent 1,000 types in English. Thus a more comprehensive list is still needed. However, a number of research questions remain unanswered.

Read (2001) stated that, in regards to vocabulary assessment, "more consideration should be given to the role of multi-word lexical items in language use" (p. 15). However, there is a lack of data and inconsistency in the testing of such fluency. Gyllstad (2007) stated that "collocation testing had thus far been conducted in a somewhat unsystematic fashion" (p. 14). Years later, Durrant (2014) still agreed, stating that "tests of second language learners' knowledge of collocation have lacked a principled strategy for item selection" (p. 443). A number of researchers have tested collocational fluency over the years, but the scope and methodologies in these studies vary greatly. First, the n-size in most of these studies was quite small, as Al-Zahrani tested 81 students, Bahns and Eldaw (1993) tested 58, Barfield (2003) tested 93, Biskup (1992) tested 62, Bonk (2000) tested 98, Farghal and Obiedat (1995) tested 34 with one test and 23 with a different test, Gyllstad (2007) tested 97, and Mochizuki (2002) tested 54. There are some exceptions. Abdul-Fattah (2001) tested 340 students, Gitsaki (1996) tested 275 students, Gyllstad tested 188 students, Jaen (2009) tested 311 students, and Koya (2005) tested 130 students. However, even these numbers are significantly less in comparison to the 549 students tested in this current study. In addition, in some of these studies the number of items tested were small. Bahns and Eldaw (1993) only tested with 15 questions while Farghal and Obiedat (1995) only tested with 22 questions. Gyllstad (2007) considered 50 questions to be a large amount of items and was the amount of questions chosen for this study for practical reasons of time limitations in regards to access to students.

Moreover, test item selection criteria varies greatly, with some having no valid scientific rationale whatsoever. For instance, Fargal and Obiedat (1995) simply chose "22 collocations of topics such as food, clothes, and weather" (p. 319) without any further explanation or

justification for these choices. Read (2007) remarks that corpora can provide "the basis for more accurate word lists from which target words can be sampled", for not only teaching but also for testing. Durrant's (2014) study confirmed this, finding that corpus "frequency data should be used as part of the process of sampling colocations for selective testing" (p. 479). Mochizuki's (2002) selection process was much better being that it was at least based on corpus data, but in his study he only chose nouns, verbs, and adjectives as pivot words, excluding adverbs. In addition, collocations were chosen from *Collins COBUILD English Collocations,* a resource derived from the *Bank of English* corpus which is mostly from written sources (ideally, a corpus should also have a balance of spoken content as well).

Bonk's (2000) selections were also restricted in that his study only examined verb-object, verb-preposition, and figurative use of verb phrases. Thus, his study was not testing general collocational knowledge but rather specific types of grammatical collocation formulaic knowledge. Furthermore, his study proactively made an attempt to use verbs in various configurations, such as past and present tenses, gerunds, and plain forms, and also consciously presented items in affirmative and negative sentences rather than relying on frequency and using how those forms occur naturally. Bonk (2000) stated this methodology aimed to "tap into learners' more complete knowledge of these forms, rather than merely their memorized knowledge of unanalyzed chunks" (p. 15). However, such an approach is not testing common collocations as they naturally occur, but rather forcefully examining all ways in which they occur.

Studies varied as well in the types of knowledge they examined. Some tested receptive knowledge (Gyllstad, 2007) while other tested productive knowledge (Bonk, 2000). Some utilized translation while others relied upon recall or recognition. The cloze test was the most commonly used testing method. Al-Zahrani (1998), Bahns and Eldaw (1993), Bonk (2000), and Gitsaki (1996) all utilized the cloze testing method. Bonk (2000) study's data found using a cloze style test to measure collocational fluency to be "relatively reliable" (p. 34), and since this study aimed to test productive knowledge, it was chosen as the testing method.

Grappling with large amounts of data

As mentioned earlier, the sheer number of collocations that exists poses as a barrier to obtaining fluency in them as well as doing research on them. Among the research done so far, Kjellmer (1987) represents the most comprehensive study, examining co-occurrence of lexi as low as two occurrences per million tokens. However, a large quantity of collocations deemed useful by native speaker intuition occur much less frequently than twice per million tokens. For example, items occurring as low as once per hundred thousand tokens can be considered worthy of teaching, such as the following lemma pairs: *nice/vacation*, *finish/workout*, and *tend/exaggerate* (Davies, 2008).

Both Hill (2000) and Pawley and Syder (1983) believed that the number of 'lexicalized sentence stems' that native speakers have at their disposal is in the hundreds of thousands. Sinclair's (1995) *COBUILD English collocations* lists 140,000 different collocations. Bahns (1993) gave a lower estimate, in the 'tens of thousands', but still referred to this as an obstacle. Hill (2000), while admitting that estimates vary, remarked that "70% of everything we say, hear, read, or write is to be found in some form of fixed expression" (p. 53). He stated that we need to accept that the 'learning load' to become fluent in a second language is not 40,000 items, but closer to 400,000 items or more. Other researchers agree, pointing out that Nation's (1990) previous estimate of undergraduate native speakers' vocabulary sizes of 20,000 'items' may be misleading, and that this only constitutes "the rudimentary base of the native speaker's lexicon" (Conzett, 2000, p. 75). Thus the sheer number is a challenge for the learner. In addition, in regards to why there is a lack of particular kind of collocational research (specifically identifying 'useful' collocations), the above statements regarding the large quantity of items to examine is clearly a barrier. There are simply so many collocations that it is difficult for one or even a number of researchers to handle.

On the criteria for identifying useful collocations

Because of the sheer number of collocations and the variety of learning burdens among them mentioned above, researchers should attempt to use criteria to identify specific collocations which are not only useful for learning but which also cause them difficulty. However, in addition to the lack of a comprehensive list, many questions remain as to which criteria should be utilized to create one. In addition to semantic transparency discussed earlier, the following

criteria have the potential to isolate and identify specific collocations that are of value for learners.

Frequency data

There are various ways to identify useful collocations. The simplest and most common involves frequency data from a corpus (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Hoey 1991; Shin 2006). While setting a frequency cut-off is unavoidably "arbitrary" (Nation, 2001a, p. 180), for teaching, a cut-off must be set in regards to the practical limitation of how many items can be directly taught during limited classroom time. As mentioned above, a large range of different frequency cut-offs have been used in collocation research, and there is still a lack of consensus on which is ideal. Biber, Conrad, and Cortes (2004) set their cut-off at 40 occurrences per million tokens, Cortes (2002) at 20, Biber, et al. (1999) at ten, Kjellmer (1990) at four, Shin (2006) at three, and Kjellmer (1987) and Liu (2003) at two occurrences per million tokens. In Moon (1994), 70 percent of the MWUs examined occur less than once per million tokens. But questions still remain as to how low a frequency cut-off can go and still contain mostly useful collocations. This is addressed by this current study's research question 1.

On statistical measures of association

Researchers have also utilized statistical measures of association, such as how Lorenz (1999) utilized mutual information data (M.I.), to identify high exclusive co-occurrence. However, M.I can be problematic. Durrant (2014) actually found the measure to not correlate with learner knowledge. M.I. emphasizes collocations whose components co-occur very often but may not have high frequencies. A good example would be *crux/matter*. In the Corpus of Contemporary American English (COCA) (Davies, 2008), corpus data indicates that *crux/matter* have a very high M.I. of 7.24 (the corpus states that collocates with an M.I. of 3 or above should be considered semantically bonded). This is why when a native speaker is presented with the following cloze sentence, they can easily produce the answer:

The crux of the m_____ is that our company needs to expand.

However, despite having such a high M.I., *crux/matter* actually have quite low co-occurrence frequency. The pair only occurred 94 times in the 450 million tokens of the entire COCA.

In comparison, the collocates *call/home* have a converse issue. Again, a native speaker would typically not have a problem completing the following cloze sentence:

My parents always make me call h_____ if I'm going to be late.

When M.I. alone is used to identify collocates, such an example would be excluded. *Call/home* has a very low M.I. score of only 0.42, while having 1,218 occurrences in the corpus, over ten times as many occurrences as *crux/matter*. The reason why *call* and *home* have such a low M.I. score, despite collocating, is that the words have high individual frequencies and frequency co-occur with a large variety of other words.

Kennedy (2003) utilized M.I. and highlighted its strength in identifying colligational relationships between words. For example, the word *perfectly* was found to always collocate with positive words and was likely to co-occur with adjectives which end in *able* or *ible,* such as *perfectly possible.* For such a purpose M.I. is useful. However, this study has a different aim in that the goal is to identify the exact MWUs most representative of high-frequency collocations. As stated above, utilizing M.I. is problematic for such a purpose because the relative word frequency can lead to less common collocates having higher M.I. scores. For instance, if data from the COCA is examined in regards to the word *perfectly,* it becomes clear that M.I. fails to identify high-frequency collocations. The highest M.I. score (11.86) identifies mere noise in the corpus, such as *perfectly/paisley-esque,* and many other examples of noise at only one occurrence in the entire corpus. However, even at ten times that frequency cut-off (a frequency cut-off of ten occurrences minimum in the entire corpus), the results using M.I. scores are still clearly not as useful as raw frequency is in identifying high-frequency collocations (see Table 8 below). Utilizing M.I. as a measure clearly identifies much less frequent collocations for all top five items, while each item identified using frequency are not only deemed useful for all practical purposes using native speaker intuition, but all also have lower M.I. scores than even the lowest ranked item in the examples given when M.I. is utilized.

Table 8

A comparison of the top five results for collocation searches in the COCA for the word 'perfectly' utilizing both M.I. and raw frequency at a frequency cut-off of ten occurrences in the corpus

Collocate	M.I.	Freq.	Collocate	M.I.	Freq.
coifed	10.07	15	*normal*	5.41	469
coiffed	9.62	49	*clear*	4.04	455
proportioned	9.18	44	*fine*	4.57	418
manicured	7.87	62	*fit*	5.17	383
serviceable	7.68	23	*legal*	4.44	335

Shin (2006) found that M.I. was strongly related to frequency and has "no additional discriminating influence" (p. 59). With this in mind, and the above clear examples of how M.I. can be problematic as a criterion for identifying collocations, this study opted to not use it as a criterion.

T-scores can also be utilized to identify collocations. A t-score provides evidence as to "how certain we can be that the collocation is the result of more than the vagaries of a particular corpus" (Hunston, 2002, p. 72). In other words, it is more of a measure of the certainty of a collocation, taking frequency into account. However, this measure can be problematic in that it has a tendency to identify collocates that have a grammatical function (prepositions, pronouns, determiners, etc.) in comparison to M.I., which tends to identify collocates which have more of a lexical and meaning relationship. The typical perspective of 'collocations' are those which are more 'fixed' in their relationship, and those are what M.I. scores identify better, but as stated above, M.I. has issues as well.

Martinez (2011) writes:

> It is very difficult to determine which one would be the ideal measure for collocation analysis; we believe the researcher should take advantage of the different perspectives provided by the use of more than one measure and therefore use as much information as possible in exploring collocations. Factors such as the purpose of the research should be taken into account, and the lexicographer should decide which statistic is the most appropriate for his study depending on his definition of collocation. (p. 766)

With this in mind, this study found both the above measures of statistical analysis to be problematic for the goals it set to achieve, and thus the decision was made to not rely on either of these methodologies.

Dispersion data

Kjellmer (1984) and Nation (2001a) stated that a collocation's dispersion, or its frequency of occurrence in a range of different categories of text, is a necessary criterion for identifying useful collocations. Durrant (2014) agreed, stating that a collocation's dispersion is a worthwhile measure to incorporate into collocation research since his study found a statistically significant relationship between it and learner knowledge. Other researchers add that, when dealing with students who are studying particular topics or have specific goals, we should present collocations that occur only in a specific range (Conzett 2000; Woolard 2000). Such collocations can easily be identified when corpora provide *dispersion* data, or the distribution of frequency among genres within the corpus. Gries (2008) believes that dispersion data analysis is essential, stating that raw frequency data can be misleading in regards to a word's general importance when the dispersion of its frequency data is unbalanced.

Dispersion has been a criterion utilized in the creation of word lists, such as Nation's (2004) *BNC 3000,* and it has been referred to as clearly being an important criterion (Nation and Webb 2011). However, while some studies on identifying useful collocations have utilized dispersion data from corpora to delimit their selections of useful collocations, none have utilized it on such a large-scale as in this current study. Furthermore, many of these studies utilized much smaller corpora in comparison to this current study. One such study is Cortes (2002). Its corpus consisted of only approximately 360,000 tokens. Biber, Conrad and Cortes (2004) also

employed dispersion criteria, but their corpus consisted of only two million tokens. These are quite small in comparison to the corpus used in this thesis (450 million tokens).

Current research thus shows how dispersion data has yet to be adequately applied to identify useful collocations. This current study will thus aim to make the value of it as a criterion more salient by answering its research question 2.

Chronological data

Chronological stability over time may also be another important criterion for identifying useful collocations to study. Clearly, learners do not need to study collocations which are dated, which occurred only during a limited point in time, or that are not yet firmly established in the language. For example, the lemma *foreign* and *soviet* occurred in the COCA's (Davies, 2010) list of high frequency collocates, but the yearly breakdown of their occurrences in the COCA reveals that 88.4% of the occurrences were from 1990-94. After that, occurrences fall off to 5.6% in 1995-99, 2.7% in 2000-04, and 2.5% in 2005-2009. It is obvious that these collocates were influenced by a particular political situation (the collapse of the Soviet Union in 1991) during a specific time period.

However, to date, no research has considered this criterion in regards to useful collocation identification. Thus, the extent to which it can be applied as a useful criterion remains to be seen. Therefore, this current study will address this gap in the research with its research question 3.

L1-L2 congruency

A multitude of researchers consider L1-L2 congruency to be an important criterion to consider when selecting useful collocations to directly focus on. By choosing such items to teach, learners are given the necessary opportunity to focus on items with which they would otherwise have a high potential of making an error with. Gyllstad (2005) gives the example of how in German the English *take a photo* can be mistranslated as *make a photo* because the German way to convey this (*ein Foto machen*) using the verb *machen,* of which the English equivalent is *make*. Zughoul (1991) gives another example of L1 interference from Arabic in that his students produced the following unnatural sentence: *the weather is kind in that country*.

Without such instruction, learners will typically directly translate from their L1 and thus produce non-native like formulations. In addition, identifying L1-L2 congruency can also improve upon the efficacy of learning. While it is true that learners make errors for a variety of reasons, including intra-lingual ones, Fayez-Hussein (1990) did find 50 percent of collocational errors were due to L1 interference, and Shin (2006) did find that one third of the high-frequency collocations he identified were incongruent to an extent with their Korean translations. Bahns (1993) recommends to not waste time teaching L1-L2 congruent collocations. However, Biskup (1992) found that some learners will even be weary of congruent collocations. Nesselhauf (2003) agreed, stating that "congruent collocations cannot be ignored...mistakes are also made when collocations are congruent" (p. 238). Moon (1997) also agreed to an extent, giving the example of how even when MWUs are congruent in both the L1 and L2, "they are unlikely to be exact counterparts, and there may be different constraints on use" (p. 58).

Liu and Shaw's (2001) study gives a specific example of this. First, their study showed that learners produced significantly less collocations (1.6 percent) in comparison with native writers (12.1 percent). They postulate that the morphological differences between Chinese and English play a role in this. They give the example of how there is a rule of inversion in English that leads to the formulation of *film-making* from *make a film* (inversion plus *ing*). However, in Chinese there is no inversion but rather the addition of the suffix *de*. So, *paidianying* becomes *paidianyingde*. Because of such differences between languages, they stated that learners will avoid unfamiliar items, or items where there are no translations equivalents, which paves the way towards fossilization. Laufer and Eliasson (1993) agreed, reporting that L1-L2 incongruency was the best predictor of avoidance of using certain phrasal verbs.

Shin (2006) gives an example of how one L1 meaning can be represented by different forms in an L2. He also highlights how one L2 form can have multiple meanings in the learner's L1. A commonly known illustration of this is how in languages in colder climates, a significantly larger amount of words exist to describe various types of snow, while in warmer climates all of these may be simply represented by one word.

In contrast, some researchers believed that a learner's L1's effect on attaining collocational fluency is marginal (Dechert & Lennon, 1989; Lennon, 1996; Ringbom, 1998). However, Nesselhauf (2003) found the exact opposite. She stated that "the learners' L1 turns out

to have a degree of influence that goes far beyond what earlier (small-scale) studies have predicted" (p. 223).

However, there is still a lack of research in regards to the extent that L1-L2 congruency is an issue in useful collocations. For example, Shin (2006) could only examine approximately 10 percent of the English collocations in his study for congruency with Korean due to time constraints. However, his study still found that L1-L2 congruency was an important factor to consider in that one third of the items examined were incongruent. Regardless, Gitsaki (1996) stated that "syntagmatic relations are more likely to differ from language to language" (p. 3). She gives the example of how the Greek learners in her study had specific problems with collocations which contained a preposition since the Greek language has many that do not coincide with English. Therefore, it is necessary for researchers to conduct contrastive analysis on all learners' L1s in question. Nesselhauf (2005) called conducting such a study "desirable" (p. 272). Liu and Shaw (2001) also recommend such studies with the goal of producing "customized syllabi applicable to teaching L2 learners of specific mother tongues" (p. 189). This study aims to fill this gap in the research with research question 9.

Regarding the direct teaching of collocations

Researchers have recommended the direct study of collocations for some time now (Mackin, 1979; Marton, 1977). Likewise Doughty and Williams (1998), Ellis (1994) and Koya (2004) all argue that collocations should be taught directly. Newman (1988) recommended the direct memorization of collocations. Gitsaki (1996) recommended their direct teaching in class.

Although rote learning is dismissed by many as outdated, the direct teaching of certain collocations/MWUs may still be advantageous. Sokmen (1997) remarks that the anathema towards rote learning has actually led to a decrease in acquisition speed, and that now the pendulum is swinging back towards the middle for a more balanced approach. Shin (2006) agrees, stating that deliberate learning itself is not a problem, but rather a "lack of balance with other ways of learning" (p. 163). In the past, discussion of more traditional methods such as paired associate learning has mainly focused on isolated vocabulary study. A vast majority of such research has shown such explicit study to be very efficient (Avery & Baker, 1997; Hopkins & Bean, 1999; Rodriguez & Sadoski, 2000). But what of collocations? Should we teach them

directly as well? Chan and Liou (2005), Hsu (2002; 2005), Lien (2003) and Lin (2002; 2004) all found that such an approach towards teaching collocations was effective.

Foremost, teachers need to expose students to useful collocations, thus enabling students to fully acquire them. However, Nesselhauf's (2005) study reveals that exposure alone is insufficient. She argues that the direct teaching of collocations is essential for developing fluency. If encounters are left to chance, then as Wollard (2000) stated, "Learning will be extremely haphazard and inefficient" (p. 26). Lewis (2000) remarks that it may be weeks, months, or even years before students re-encounter a particular collocation. Bahns and Sibilis (1992) found a similar issue, in that in their study mere exposure had little or no effect on improving collocational fluency. Furthermore, when students are exposed or taught collocations directly, it seems as if they are being introduced unsystematically (Howarth, 1996). Gairns and Redman (1986) remarked that teachers typically just deal with collocations as they appear in materials, which is clearly inefficient and disorganized.

Gairns and Redman (1986) note that the most common way teachers deal with collocations is as they appear in the textbooks they use, and state that this is not ideal, if effective at all. Biskup (1992) explained why this is, stating that "when encountering a new collocation, a learner does not make a conscious effort to understand or memorize it as it poses no specific perception problem to him or her" (p. 87). Lewis (2000) and Wollard (2000) also agreed, stating that directly focusing on collocations will bring students' attention to very high frequency words that they are already familiar with but do not realize are actually occurring formulaically. Lewis (2000) agreed, stating that while he agrees "that learners should take responsibility for their own learning, they should not be taking responsibility for choosing which language items are more linguistically useful" (p. 18). Myers and Chang (2009) suggested that learners cannot simply gain collocational knowledge on their own and that they need some sort of guidance.

In particular, a number of researchers stated that L1-L2 incongruent collocations in particular should be taught directly (Bahns, 1993; Gairns & Redman, 1986). Laufer and Girsai (2008) found that students who studied using a contrastive analysis method outperformed meaning focused and non-contrastive form focused methods. They describe it as a "perfect 'pushed output' task that requires stretching one's linguistic resources" (p. 710) because it involves a higher involvement load.

Furthermore, since even advanced learners have been shown to have low collocational fluency, such students may need to learn how high-frequency vocabulary co-occurs despite having already mastered such isolated vocabulary's semantics. Lewis (2000) echoed a similar remark, stating that "some students already know a lot of 'simple' words but are not aware of what those words can do for them because they haven't noticed their common collocations" (p. 24). In connection, Woolard (2000) made an interesting statement when he said that "learning more vocabulary is not just learning new words, it is often learning familiar words in new combinations" (p. 31).

From such a perspective, the collocation itself is being considered as a 'word'. This is logical since there is evidence that shows that this feature of language is stored in the brain in chunks in the same way that isolated vocabulary items are (Ellis, 1996; Wood, 2004; Wray, 2000). Hill (2000) agreed, stating that "in the same way that we teach individual words we need to teach collocations. Rather than wait for students to meet common collocations for themselves, we need to present them in context just as we would present individual words" (p. 60). He thus suggests that every time a teacher teaches a new word, that word should be taught with its common collocate.

There is also a need for larger contextual support as well to help learners master all aspects of vocabulary depth knowledge, such as restrictions on use, etc. Woolard (2000) stated that teachers must become aware of the need to incorporate such co-textual information into their teaching. Thus, not only should learners be taught collocations directly in the form of MWUs, learners should also be given additional contextual support to help them truly master all the knowledge necessary to use the MWU properly. For instance, if *pro bono* is taught it would be ideal if a full contextual sentence accompanied it which brings attention to the fact that phrase is almost exclusively used as a legal term.

Conclusion

In conclusion, this literature review has shown that there is still a considerable amount of disagreement as to what should and what should not be considered a collocation. It was shown that there are various to define and approach collocations, with each having strong and weak points depending on the goal of the research. Out of all of the previous approaches taken toward

understanding and defining collocations, the lexical approach along with some aspects of the structural approach has been shown to be ideal for the purposes on this current study. So, for this current study's goals, the approach of using frequency of co-occurrence as the main criterion for identifying collocations was found to be ideal. Research indicated that criteria such as M.I. has the potential to not accurately reflect natural language to an extent, and that rather by using frequency of co-occurrence collocation identification becoming either too inclusive or exclusive can be avoided. It was also shown that identifying collocations by counting words as lemma was preferable to types or word families.

The literature review has also shown that if a study is to be second language learner-focused, then rigid classification and definition of collocations as distinct and separate from literals is not ideal because of factors such as L1-L2 congruency. In regards to word counting and MWU identification, concgramming was shown to be the best approach to count co-occurrence in a way that best reflects natural language. Seven word long MWUs which utilize native speaker intuition was also found to be the best option to provide learners with the most useful information on how collocates are typically used. The value of consideration for colligation was also highlighted.

In general, previous research views collocational knowledge to be a priority and of significant importance for obtaining fluency in a second language. They are a major part of any language and occur ubiquitously in any text, and some researchers believe that the vast majority of the language we speak actually occurs in chunks. Some researchers actually view collocation as the central organizer of language itself.

Previous research indicated that collocational knowledge is a highly valuable skill for L2 learners to master, and thus worth focusing on. Researchers have found that collocational fluency helps learners read quicker and make grammatical judgments quicker and more accurately. They also found that output was smoother and more fluent. So, in comparison to learning isolated vocabulary and/or decoding or encoding information word by word, being aware of collocational relationships between words should enable learners to process language more efficiently. Furthermore, the concept of mnemonic hooks highlighted how learning MWUs can actually be easier in comparison to learning isolated vocabulary items. It is therefore logical

to study vocabulary via MWUs instead of isolated vocabulary since they have a lower learning burden.

However, researchers also revealed that there is a severe lack of collocational fluency among second language learners from a wide variety of backgrounds. Learners from Asia, the Middle East, and Europe were all shown to lack collocational fluency, even those who had advanced fluency. In fact, collocational knowledge was shown to not improve as learners improved upon other aspects of fluency. Many researchers indicted that the vast majority of errors learners made were actually collocational in nature. Multiple researchers found that approximately 50 percent of the collocations learners produce were either deviant or totally wrong.

If so, many learners are lacking in collocational fluency, then it is only logical to investigate why this is. A number of researchers point out a variety of issues that lead to a high learning burden for collocations. First, the very large quantity of collocations a learner must master is a major barrier. Some researchers believe that there are hundreds of thousands of collocations that exist. Other researchers point out that the way collocations are formulated is also very complex, and often arbitrary. Thus, it is very easy for learners to make errors by over-generalizing with or by underusing certain collocations.

Another aspect of collocations that make them difficult to learn was shown to be their semantic transparency. Some researchers found that semantically restricted collocations were found to be more difficult to learn, while others also indicated that learners can struggle with literal collocations as well. A number of researchers pinpointed specific types of collocations that learners from specific backgrounds struggled with. For instance, Taiwanese students were found to especially struggle with verb-noun collocations, and when they did the vast majority of the errors was with the verb.

The next step to examining why learners lack collocational fluency is to examine whether or not there is a lack of research and resources, and/or a lack of focus in second language instruction. In fact, previous research indicated that this is the issue. Collocations were found to not actually be taught directly to students although a number of researchers believed that this would be ideal. Teachers have been aware of the value of collocations and that they should be

taught directly for some time now, but they still are not. This occurs because there is still a lack of research and resources to tap for such instruction.

Previous research has limited the scope of what was examined so severely that to date, no comprehensive resource exists. This stems from the very large amounts of data that need to be analyzed to produce such a resource, in addition to a lack of consensus on how such an analysis should be conducted.

In regards to the criteria that should be used in such an analysis, previous research also indicated that there is a lack of consensus and/or research itself. However, many researchers agree that frequency of co-occurrence is an important criterion to consider. Because of the hundreds of thousands of collocations that exist, setting a frequency cut-off is, while unavoidably arbitrary, necessary. Which cut-off is best is a question that remains unanswered, though.

In addition, as mentioned earlier, the semantic transparency of a collocation is also another important criterion to consider. However, it remains to be seen what percentage of high-frequency collocations are semantically transparent or not. In addition, as also mentioned earlier, the issue of L1-L2 congruency can make the criterion of semantic transparency moot. Another criterion mentioned above that can also be problematic is M.I. Previous researchers have not only found it to be problematic, they also found it to have no discriminating influence in comparison to frequency. Using dispersion data is another criterion that many researchers agree is of importance. Such data was found to be useful in identifying collocations that occur across a wide variety of language, thus helping to identify only truly useful items. Furthermore, using dispersion data in addition to frequency data helps avoid issues of being misled by raw frequencies of items which only occur in one type of language, such as academic language. Chronological data was also shown to be an important criterion to consider, but this literature review has shown that research is totally lacking in regards to its usefulness as a criterion. Thus, this current study will aim to fill this large gap in the research. Another very important criterion mentioned already was L1-L2 congruency. A wide variety of researchers agreed that this is an important criterion to consider when the aim is to identify collocations which have a higher learning burden for learners, and thus need to have additional focus.

There are too many researchers who feel collocations are valuable for second language learners to list, but some recent examples include Almela and Sanchez (2007), Chon and Shin

(2009), Hoey (2005), Liu and Shaw (2001), Nesselhauf and Tschichold (2002), and Webb and Kagimoto (2011). There are similarly too many researchers who agree that collocations are should be taught directly to list, but some recent examples include Bonk (2000), Chan and Liou (2005), Hsu (2002), Koya (2004), Keshavarz and Salimi (2007), and Myers and Chang (2009). However, there is still shown to be a lack of research and resources in a number of areas and a number of questions remain as to how to conduct such research and/or create resources that can help learners acquire collocational fluency. This current study aims to fill these major gaps in the research by answering the following research questions with the ultimate goal of validating a methodology to identify useful collocations and creating a resource that teachers and learners can begin to use for the direct study of collocations.

Chapter 3

Research Methods and Techniques

Introduction

This section of the thesis will discuss the overarching approach to the proposed research, the data set, instruments used, what type of data will be collected, and how it will be analyzed. The ethics and politics concerned with this study will also be discussed.

Research Paradigm

The overarching approach to the proposed research is post-positivist, in that the nature of the research must employ measures that approximate reality, while admittedly possessing weaknesses that are unavoidable. For instance, it is impossible to choose a high-frequency count cut-off and show that any occurrence below that particular number is a low-frequency item. Nation (2001) stated that setting such a frequency cut-off is unavoidably arbitrary but necessary for practical reasons of delimiting what should be taught directly. Furthermore, a post-positivist approach also operates under the assumption that one singular answer is attainable, and is not preoccupied by multiple perspectives. It also avoids the pitfalls of breaking down collocations into more and more restricted categories, and instead aims to find one answer: what are the most frequent, useful collocations. Such an objective approach, which employs quantitative analysis with the aim of discovering the best approximation of reality, is ideal for such a study. Such an

approach frees itself from the inadequacies of black and white thinking by admitting that results will never be unequivocal, but will rather be the best approximation possible within unavoidable constraints.

More specifically, this study will be a corpus-driven exploration of high frequency MWUs. It aims to first identify necessary criteria for inclusion, and use these criteria to find the most useful collocational lemma pairs. This study also aims to devise a methodology to find the most frequent MWU that these lemma occur in to help present them to language students in the most useful contexts. Using lemma instead of word types, as Shin (2006) did, is unique compared to previous research. Combining lemma data helps to consolidate data. So, MWUs such as *take a walk, took a walk, taking walks* will all be listed and compared as if one 'item'. Listing them separately (using types) and presenting them to students at separate times (due to frequency differences) would make an already copious amount of data even more excessive, and also make for less efficient learning since learning such units at the same time is clearly ideal. Semantic transparency and contrastive analysis will also be conducted to help identify items students need to spend the most time on, and to reduce the overall learning load. Finally, this study will test Japanese university students' knowledge of a balanced sample of these items.

Just as many models of language are complementary, rather than any singular model being a definitive paradigm, this study will draw from more than one approach to examine all pertinent aspects regarding collocation. This study will draw from the structural approach only when it is appropriate to deal with MWUs whose counts are strongly affected by colligation, while mainly being driven by the lexical approach's tenet of the importance of raising students' awareness of lexis and the way words combine to help them attain fluency.

Data source

One of the most useful resources for identifying common collocations/MWUs is corpora (Meijs, 1992; Noel, 1992; Francis, 1993). But which corpus should be used for the identification of useful collocations? Shin (2006) stated that a large corpus with a large variety of texts is essential for producing data which best resembles natural language. Thus Kjellmer's (1994) use of the 1-million-token *Brown Corpus* (Nelson and Kucera, 1979) may not have produced the data which truly reflects natural language, despite it being one of the largest corpora at the time.

Through computer technology larger and larger corpora have been compiled. In recent years, many researchers have relied on the 100- million-token *British National Corpus*, or *BNC*, for collocational research (e.g., Durrant & Schmitt, 2009; Shin, 2006). However, the *BNC* stopped being developed in 1993 and has been referred to as being past its sell-by date (Kilgarriff, Atkins & Rundell; 2007). In fact, Durrant's (2014) study showed that Davies' (2008) *Corpus of Contemporary American English*, or *COCA*, was "more strongly related to learner knowledge than the older and smaller *BNC*" (p. 472). Thus, the *COCA* can be considered a better choice as it is four times larger than the *BNC*, and it is still being added to today. It consists of American English, which is the variety of English that the target learners in this current study learn. Furthermore, it has a wide and balanced dispersion in regards to genres and spoken versus written content.

Specifically, the COCA is broken down into equal genre sections called *spoken, fiction, popular magazines, newspapers,* and *academic journals.* Its spoken section currently contains approximately 109 million tokens, which are transcripts from more than 150 different television and radio programs. Its fiction section contains approximately 105 million tokens, which consist of short stories and plays from literary magazines, children's magazines, popular magazines, first chapters of first editions books from 1990 to the present, and movie scripts[6]. Its magazine section contains approximately 110 million tokens, sourced from nearly 100 different magazines with a balanced mixture between specific domains (news, health, home and gardening, women, financial, religion, sports, etc.). Its newspaper section contains approximately 106 million tokens, which are from ten newspaper across the U.S., including *USA Today, New York Times, Atlanta Journal Constitution, San Francisco Chronicle,* among others. In general, there is a good mixture of different sections of the newspaper as well, such as local news, opinion, sports, and financial sections. Finally, the corpus' academic section consists of approximately 103 million tokens from nearly 100 different peer-reviewed journals selected to cover the entire range of the Library of Congress' classification system, such as having equal percentages overall

[6] It should be noted that ideally movie scripts should have been included in the spoken section, but this did not have an impact on this current study. However, corpus compilers should take note of this.

and by number of words per year of the Library of Congress sections, such as *B* (philosophy, psychology, religion), *D* (world history), *K* (education), *T* (technology), and so on.

This study will examine lemma collocates of Davies' (2010) most frequent 5,000 lemma list (see Appendix 1) that occur at a span of either four words to the left or right of the pivot word derived from the entire COCA (Davies, 2008) of nearly 800,000 collocations (available commercially). Not only is this span range the exact maximum that the COCA's interface can process collocational searches with, it is also the span recommended by Jones and Sinclair (1974). This list is available for purchase from http://www.collocates.info/purchase.asp. This list was then delimited by a frequency cut-off of approximately one occurrence per million tokens (500 occurrences in the corpus) and only content words (nouns, verbs, adjectives, and adverbs) were included in the analysis. These two criteria resulted in a list consisting of 25,969 lemma pairs (see Appendix 2). The *COCA* (Davies, 2008) corpus itself will be utilized to collect example sentences for these pairs to aid in the identification of MWUs most representative of the lemma pairs. The corpus will also be used to collect chronological and genre dispersion data to aid in identifying only items with balanced data dispersion.

Instruments

The *COCA's* (Davies, 2008) online interface will be utilized to search for chronological and genre dispersion data and example sentences for each collocational lemma pair. Anthony's (2013) *AntWordPairs,* custom software written specifically for this current research project, will then be used to examine 500 examples sentences which each pair occur in to extract the most frequent MWU most exemplary of those lemma. This method will reveal the most common MWU two lemma occur in, such as how *take a walk* is the exemplar of the lemma *take* and *walk,* as discussed above. From two to seven-word MWUs will be searched for, and only MWUs which occur in more than five percent of the total example sentences examined will be included.

Since this methodology is completely new, there is no previous precedent for a cut-off percentage, such as the five percent cut-off mentioned above. However, when the data is examined it is clear that this cut-off provides robust enough data to accomplish this study's goal while removing unnecessary data, which also eases the processing load on the computer. Processing the data was extremely heavy and required the software to run for hours on end, and

software crashes resulted when the load was below five percent. Thus, by utilizing some kind of practical cut-off that still provided the necessary data to accomplish the task, actual processing and analyzation of the data becomes possible. Keeping all data which occurred at either five percent or more of the constructed corpora was found to provide enough data to help identify the most common MWU and to extend that MWU beyond the top identified item when the native speaker deemed necessary. For example, this is clear when the data for the lemma *able* (adjective) and *afford* (verb) is examined (see Table 9 below). For this lemma pair, *be able to afford* was chosen to be the MWU most representative of how the lemma *able* and *afford* occur together since it is not only an extension of the top identified MWU (*able to afford*) which occurred in 97 percent of the example sentences examined, but it also exhibited a high percentage of total occurrences (62 percent) as well. Beyond that, the percentages drop off significantly to 12.4 percent and below and the next extension of *able to afford* occurred in only 11.6 percent of the total. Therefore, using this data and native-English speaker intuition it was deemed appropriate to stop extending at that point and choose *be able to afford* as the MWU to represent the lemma pair.

Table 9

MWUs identified via concordance software processing of a corpora of 500 example sentences in which the lemma 'able' (adjective) and 'afford' (verb) both occur at a limit of five percent or more of the total sentences

Occurrences Out of 500 Example Sentences	Percentage of total	MWU identified
485	97	*able to afford*
310	62	*be able to afford*
62	12.4	*able to afford to*
58	11.6	*not be able to afford*

54	10.8	*able to afford the*
54	10.8	*to be able to afford*
53	10.6	*able to afford a*
53	10.6	*'t be able to afford*
52	10.4	*able to afford* [subject pronoun]
50	10	*be able to afford to*
50	10	*been able to afford*
36	7.2	*won't be able to afford*
36	7.2	*be able to afford* [subject pronoun]
33	6.6	*be able to afford a*
32	6.4	*may not be able to afford*
32	6.4	*be able to afford the*
29	5.8	*will be able to afford*
26	5.2	*being able to afford*

All items will also be examined for colligational issues. Data will first be examined to determine what kind of colligational issues exist, such as how it may be beneficial to replace items such as days of the week, ordinal and cardinal numbers, etc., with a unifying marker to allow their occurrences to be counted together. Then, *GoTagger* (Goto, 2005) and *Textcrawler* (2011) will be used together to first adjust for homonyms, and then *Textcrawler* will be used to adjust such items.

Data Collection

First, duplicate entries will be manually removed from items delimited by part of speech and frequency in Davies' (2010) list. Duplicates occur when a collocate also happens to be a high frequency lemma. For example, this study began with a list of the most common collocates of the top 5,000 lemma in the COCA. Within this top 5,000 lemma, both *sing* and *song* occur. The list of collocates uses each as a pivot word. Thus, at one point in the list *sing/song* will be identified as a common collocate, but at another point *song/sing*. However, both will result in

the same exact data and a duplicate MWU being identified. Since the goal of this research is to produce a resource for learners to directly study, having duplicate entries is clearly unacceptable and thus such duplicates were searched for and removed.

Then, chronological and dispersion data for the remaining pairs using the COCA's online interface will be collected, and only items deemed to have balanced dispersion of data will be kept. A variety of parameters will be experimented with to determine if the corpus data and methodology used is deemed by a native English speaker who is an experienced English teacher sufficient enough to identify items with balanced dispersion, and if not, that native speaker's intuition will be relied upon to manually make judgments as to whether an item has value for learners of general English. 500 example sentences for each pair from the COCA will then be collected, which will then be used for collocational exemplar searches using *AntWordPairs* (Anthony, 2013).

Finally, data regarding Japanese university freshmen's (Kansai Gaikokugo University) knowledge of these collocational exemplars will be collected. An attempt will be made to test as many students as possible from the same university. An attempt will also be made to test students from as wide of a range of proficiencies possible, and their proficiencies will be confirmed by the collection of recent TOEFL score data. Ideally, data would be collected for sophomore, junior and senior students as well but access to such students was not possible at the university. Since it is well-known that Japanese students have far superior visual recognition versus aural recognition skills in English (Hyland, 1994; Kaneko, 2008) because of the focus in tertiary and secondary Japanese schooling on explaining grammar points about English in Japanese and high-stakes tests being mostly written tests, the proposed test will be a visual diagnostic test to determine the highest possible familiarity with the items. Productive skill will be measured by taking a direct approach, and thus students will be tested with productive cloze questions. The rationale behind testing production instead of receptive knowledge was because if the results showed high productive knowledge, then receptive knowledge could be assumed and further testing would not be necessary. However, if results showed low productive knowledge, then clearly more research would then be called for to determine receptive knowledge. If receptive knowledge was tested first, then regardless of the results, another study examining productive knowledge would be required. The test will contain 50 questions. The

final list of MWUs identified will be broken into five separate frequency ranges by the total number of items, and an equal amount of items from each of these five sections will be selected according to L1-L2 congruency ratings. Since the aim was to select 50 questions and L1-L2 congruency ratings will be given from 0-12, items were congruency ratings of 3, 6 and 9 were excluded from selection to maintain an equal selection. In each of the five frequency sections, the item with the lowest frequency count in its section for each of the ten L1-L2 congruency ratings was chosen. Thus, the item with the lowest frequency count in each section that received a score of 0 for L1-L2 congruency was chosen. Then, the item with the lowest frequency count in each section that received a score of 1 for L1-L2 congruency was chosen, and so on for L1-L2 ratings of 2,4,5,7,8,10,11 and 12. Test questions will thus be balanced in regards to item frequency and L1-L2 congruency to help determine whether these two factors show any correlation with student knowledge.

This test will also be discrete point, making every effort to only test collocational fluency, and not any other language skills. Specifically, no supporting context words in the cloze sentences will be beyond the most frequent 3,000 word families of English (*COCA* and *BNC* corpus data combined) (Cobb, 2013). Furthermore, only the least frequent lemma in the lemma pair will be the target item to answer. For instance, if the MWU in question is *get upset* and the more common lemma (*get*) was chosen as the target (e.g., *I hate it when you g _ _ upset like that and starts screaming*), then learners unfamiliar with the MWU itself but familiar with the common word *get* may be able to guess the answer. Thus, this step avoids such a problem. In addition, the first letter of the target item will be provided to avoid other answers. When this does not suffice, more letters will be provided. Determining the need for this will be done by validating the test with native English speakers to ensure that the example sentence is sufficient in prompting the correct collocate.

Data analysis

First, the MWUs identified will be rated for semantic transparency. Grant and Bauer`s (2004) taxonomy, which breaks down MWUs into *literals*, *figuratives, ONCEs,* and *core idioms,* will be utilized to judge semantic transparency. Items which do not fall into the above categories will also be marked as such.

Then, an example sentence will be written for each MWU, and both the MWU and the sentence will be translated into Japanese by five volunteers who are professional Japanese translators with native-like ability in English, two of which are also university teachers of English as a second language in Japan. Next, contrastive analysis will be conducted to rate the L1-L2 congruency of the MWUs and their Japanese translations by a volunteer who is a professional translator and also a university teacher of English as a second language in Japan. Due to the extremely large and time consuming task of rating L1-L2 congruency (over 150,000 English words and their translations to examine), only one qualified person could be found to complete the task. It is clear that multiple rating would have been ideal, but this simply was not feasible due to the difficulty of finding other qualified volunteers for such a task. Therefore, it is acknowledged that more research will still need to be done to further valid this study's results.

Next, frequency data will be examined to ensure that the resulting MWUs and their example sentences do not constitute a learning burden that is not practical.

Students (N=549) will then be tested on their knowledge of these MWUs, and statistical analysis will be conducted to determine whether or not student proficiency level, MWU frequency, or L1-L2 congruency play a role in determining student knowledge. Specifically, multiple regression analysis with student TOEFL scores as the dependent variable and item frequency and L1-L2 congruency as independent variables will be utilized to determine if any correlation exists between these factors.

Ethics and politics

Since the majority of the study involves the collection and analyzing of data from the COCA, any copyright issues involved with its usage will be considered and adhered to. The sheer size of the proposed data collection and necessity for translation has necessitated a team of research assistants to be assembled. In addition to the above described volunteer translators, a team of five native English speaking teachers of English in Japan was assembled to collect and analyze data, three of which are university professors and two of which are junior high school teachers. The contributions of these volunteers to the research is simply data collection and translation, which will then be analyzed and discussed. Research assistants were not utilized for

reviewing literature, research writing, or any other responsibilities typically held by a PhD candidate whatsoever.

Students tested at Kansai Gaikokugo University were informed of the proposed research and any rules set forth by the university and/or the Japanese government regarding research was respected. Data was anonymised and only utilized for research purposes, and student consent for this was obtained. Participation was voluntary, but since it was beneficial to students as they are studying English at a foreign language university, learning such items is part of the goals of their course's curriculum. Furthermore, discovering their strengths and weaknesses regarding collocations will help them to develop their fluency, and thus participation was considered worthwhile and ethical.

Conclusion

This section of the thesis explained this study's research methods and techniques. The study will employ a post-positivist approach because of the necessity to employ measures which approximate reality. A search for collocates with the top 5,000 lemma of the COCA will be conducted, which will also be used for concordance searches, and frequency, dispersion and chronological data collection. Custom software will be utilized to identify MWUs. Duplicate entries will be removed, and parameters will be set in regards to frequency, dispersion, and chronological data. Semantic transparency analysis will be conducted on all MWUs. All MWUs will also be translated and then an L1-L2 congruency analysis will be conducted. A balanced sample of the data will then be used to create a test in which Japanese university students' knowledge of the items will be determined with.

This chapter also discussed the study's ethics and politics. It explained that any copyright issues would be adhered to in regards to the corpus data. It also explained that research assistants would be utilized to help collect and analyze data, but that these assistants would not participate in reviewing literature, research writing, or any other responsibilities typically held by a PhD candidate. In addition, university and governmental rules in regards to the students being tested in this study, and any ethical issues in regards to their participation was also explained.

Chapter 4

Answering the Research Questions

Introduction

This chapter will first discuss the scope of the research questions. It will then list the procedures, results, discussions and conclusions for all nine research questions. Finally, it will conclude with a research questions and answers summary section.

Scope of the research questions

Previous research indicates a lack of a large-scale resource which identifies high-frequency collocations that are worthy of direct instruction or study. By examining the reasons why this gap in the research exists and how it could be solved, it becomes quite clear why this gap exists.

First, a methodology had yet been developed that utilizes the concgramming method and takes into consideration all the criteria in this study, especially at such a large-scale, and thus a significant amount of time in this current study was spent creating and testing such a methodology. Some methods used proved fruitful, while others did not. Since some were extremely complex and time consuming, the results should be helpful for future researchers to avoid spending time taking steps which are not worthwhile. Furthermore, as each step was taken towards finally identifying the collocations and testing them, a number of new discoveries were made requiring the rethinking and planning of this research project's approach. For instance, to accomplish one particular task, software did not even exist and an expert in the field had to be relied upon to create complex custom software specifically for this study.

Second, the sheer amount of data that needed to be analyzed was staggering. No one researcher could accomplish such a task. This necessitated the creation of a research team consisting of volunteer data collectors and translators. For instance, any collocation worthy of direct study should have balanced dispersion, but such data was not readily available in an easy to analyze form from the corpus used in this study. Therefore, such data had to be copied and pasted from a website manually for over 10,000 items. Another example of why such a team of volunteers was necessary was how not all collocations have an equal learning burden. For instance, when a collocation is said in a very different way in a learner's mother tongue that item

will have a much higher learning burden (L1-L2 incongruency). So, if a specific group of learners is to be tested (in this case, Japanese university students) on their knowledge of the collocations identified in this study, a balanced selection of items needs to be taken with such issues as L1-L2 congruency in mind. The only way to achieve this is for each collocation to be translated into the L1 in question, compared, and be given a rating in regards to its congruency. This created the necessity to translate and compare over 10,000 items, a task that could only be done by a number of volunteers (in this study's case, five Japanese native speakers with native-like fluency in English).

Throughout this thesis, not only corpus data will be relied upon to answer the research questions, but native speaker intuition will be as well. Inherently, native speaker intuition is subjective and flawed to an extent, but as this study progresses, it will be revealed that the usage of it is necessitated and in fact improves upon results. Certain questions this study put forth cannot be answered with technology alone and native speaker intuition must be utilized (research question five), or can be answered with technology but would result in data which experienced native speaking teachers deem unacceptable than would be when native speaker intuition itself is relied upon (research questions two and three). Since native speaker intuition can be subjective, only native English speakers with over ten years' experience teaching English to the target group of learners (Japanese students) were utilized in this study. Such teachers can rely upon their education and teaching experience to make judgments on the appropriateness of what is worth teaching and what is not. They were instructed to make judgments with actual students in mind, and curriculum design, considering whether or not they would truly include items deemed worthy of instruction in the actual courses they teach to Japanese learners. The ability of the native speakers who participated in this study to achieve the tasks they were assigned is evident in the results of research question eight. Unfortunately, due to the practical constraints of the extremely time consuming work, it was not possible to utilize multiple native English speaker judgments for all of the experiments. However, within these constraints, the results of this current study still proved fruitful, albeit with limitations to how its findings can be interpreted.

With these above issues in mind, it became evident that the goal of this study was an ambitious one with practical limitations that prevented it from being accomplished in the past. However, as stated above, a number of steps were taken to overcome these obstacles. With the

help of volunteers to collect data, translators, and a software engineer, the goal was achieved to the best extent possible within the constraints of a single dissertation and practical limitations to manpower and technology.

RQ1 : What is a frequency data cut-off for lemmatized concgrams that results in a list consisting of 2-3,000 word families?

Introduction

Because of their large numbers, determining a frequency cut-off is a necessary step in identifying the most useful MWUs to directly teach or study. Biber, Conrad, and Cortes (2004) set a self-admittedly conservative cut-off at 40 occurrences per million. Cortes (2002) limited examined items to 20 occurrences per million tokens, Biber, et al. (1999) considered up to ten occurrences, Shin (2006) examined as low as three occurrences, and Kjellmer (1987) collected data for items occurring two times per million. But questions still remain as to how low a frequency cut-off can go and still contain mostly useful collocations. Thus, the following experiment will determine the most ideal corpus frequency data cut-off for identifying MWUs most representative of high-frequency lemmatized concgrams.

Corpus size and cut-off frequency are very important aspects of language analyzation such as this study. For instance, with a corpus size (425 million words) such as that used in this study, different frequency cut-offs will result in very different quality results. This study aims to identify collocations which would help learners of general English. A cut-off of three occurrences per 425 million tokens (the lowest cut-off in the list) in Davies' (2010) list of collocations from the corpus would not be ideal because this results in lemma pairs being identified which have little value to learners, whose co-occurrence seems by native speaker intuition to be at random, such as *entertain/adjourn*. However, as the frequency cut-off moves upward further, more and more items are identified which exhibit meaningful relationships that would be useful for learners of general English. For instance, this study experimented with a variety of cut-offs and determined that 500 occurrences per 425 million words was ideal. At this cut-off, the last five lemma pairs identified in the list were *steal/try, spend/study, would/satisfy, widely/accept,* and *home/fly*. When relying on native speaker intuition, these items identified are

deemed as being language commonly found in general English. These examples show the relationship between corpus size and frequency, and the importance of choosing an appropriate frequency cut-off.

Materials

In this experiment, the source for collocational lemma pairs was Davies' (2010) *word list plus collocates*, which consists of 739,254 lemma congrams (see Appendix 2). It was compiled using frequency data from the 450 million token *COCA* that was tagged with the *CLAWS 7* part of speech tag set (University Centre for Computer Corpus Research on Language, n.d.) and only includes collocates with three or more occurrences. It consists of the most frequent lemma pairs that co-occur with the most frequent 5,000 lemmas in the corpus.

Cobb's (2013) program *Vocabprofile* was used to count how many word families the collocations consisted of. This was to ensure that the number of word families did not exceed 3,000 word families, what is considered to be high-frequency (Nation, 2001b), since this study aims to identify collocations that would be useful for learners of general English.

Procedure

Davies' (2010) collocation list was utilized as a starting point and a frequency cut-off was set. Nation (2001a) suggests 2,000 word families as "practical and feasible" (p.96) in regards to direct teaching, while Nation (2001b) suggests a limit of 3,000 word families. Thus assuming the collocations selected were deemed useful, this study aimed for 2-3,000 word families.

The rationale behind this approach is the perspective that when MWUs are selected for direct instruction, these units are not only teaching about collocations and how they formulate into MWUs, but also the vocabulary in the MWUs themselves. For example, if a highly inclusive frequency cut-off was chosen for this study which resulted in the identification of many low-frequency vocabulary as collocations, those items would put undue burden on the learners since previous research indicates that only high-frequency vocabulary should be taught directly while low-frequency vocabulary should be acquired through other activities such as extensive reading. So, by aiming to only include collocations of high-frequency vocabulary which are also

mostly high-frequency vocabulary, the items identified will not result in high cost/low value for learners.

A number of frequency cut-offs were piloted to determine how many useful collocations there were at each level. The study began at the highest cut-off set by Biber, Conrad, and Cortes (2004) of 40 occurrences per million tokens, and progressed to Kjellmer's (1987) two occurrences per million. Then the 25,000 word family BNC and COCA list in the *Vocabprofile* program (Cobb, 2013) was utilized to determine how many word families the collocations consisted of to ensure that those selected did not exceed 3,000 word families.

Only content words (i.e., nouns, verbs, adjectives, adverbs) were considered. Duplicate entries were also removed, since often the collocation that occurred was a node word itself within the most frequent 5,000 lemma of Davies' (2010). The 'usefulness' of a sample of the pairs were then judged by a native speaker to ensure that the list was not overly inclusive.

Results

The cut-off of two occurrences per million tokens utilized in Davies' (2010) resulted in a list of lemma pairs consisting of only 1,874 families plus off-list types (see Appendix 3). It was thus determined that a more inclusive cut-off could be considered given the pedagogically feasible goal of teaching between 2,000 and 3,000 word families (Nation, 2001a; 2001b). Pairs occurring once per million tokens consisted of 2,789 families plus 140 off-list types (see Appendix 4), and pairs occurring once per 500,000 tokens consisted of 4,778 families[7] (see Appendix 5). Therefore, the cut-off of one occurrence per million tokens was determined to be ideal.

When the lemma pairs remaining at this cut-off point were processed with *Vocabprofile* (Cobb, 2013), it was found that these covered 83.14 percent of the top 3,000 word families (in bold in Table 10 below). Also of note is the fact that 96.74 percent (in bold in Table 10 below) of the tokens in the lemma pair list occur within the top 3,000 word families. A more detailed breakdown of the data can be seen in Table 10 below.

[7] Data set was too large to be processed via *Vocabprofile* and thus Heatley, Nation and Coxhead's (2002) *RANGE* program was utilized instead for this file. It should be noted that both programs function identically and use the same BNC/COCA combined reference data.

Table 10

Word frequency breakdown of lemma pairs occurring once per million tokens according to Vocabprofile's 25,000 word families of the BNC and COCA. ('K' represents 1,000 word families. Thus, K-1 equals 1-1,000 most frequent word families, K-2 1,001-2,000, and so on.)

Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token %
K-1 Words :	887 (32.46)	1,247 (36.00)	17,277 (68.38)	68.38
K-2 Words :	757 (27.70)	968 (27.94)	4924 (19.49)	87.87
K-3 Words :	628 (22.98)	721 (20.81)	2242 (8.87)	**96.74**
K1-3 Coverage:	(**83.14**)			
K-4 Words :	240 (8.78)	247 (7.13)	399 (1.58)	98.32
K-5 Words :	114 (4.17)	114 (3.29)	154 (0.61)	98.93
K-6 Words :	51 (1.87)	54 (1.56)	71 (0.28)	99.21
K-7 Words :	19 (0.70)	19 (0.55)	22 (0.09)	99.30
K-8 Words :	16 (0.59)	16 (0.46)	18 (0.07)	99.37
K-9 Words :	8 (0.29)	8 (0.23)	9 (0.04)	99.41
K-10 Words :	2 (0.07)	2 (0.06)	2 (0.01)	99.42
K-11 Words :	5 (0.18)	5 (0.14)	9 (0.04)	99.46
K-12 Words :	1 (0.04)	1 (0.03)	1 (0.00)	
K-13 Words :	1 (0.04)	1 (0.03)	1 (0.00)	
K-14 Words :	2 (0.07)	2 (0.06)	2 (0.01)	99.47
K-15 Words :				
K-16 Words :				
K-17 Words :	2 (0.07)	2 (0.06)	2 (0.01)	99.48
K-18 Words :				
K-19 Words :				
K-20 Words :				

K-21 Words :				
K-22 Words :				
K-23 Words :				
K-24 Words :				
K-25 Words :				
Off-List:		56 (1.62)	133 (0.53)	100.00
Total	2,733	3,464 (100)	25,266 (100)	100.00

Total word families found (2,733) plus off-list types (56): **2,789**

This one occurrence per million cut-off resulted in a list of 25,969 lemma pairs (see Appendix 2). However, many duplicate entries existed in this list because sometimes the collocate of a pivot word also happened to be a pivot word itself. For instance, the lemma pairs *indicate/clearly* and *clearly/indicate* both exist in the list. Such instances were manually checked for and 12,271 of them were found and removed. In addition, any proper nouns, noise in the data (such as the corpus' unusually high-frequency of the rare *supra/note*), and language not suitable or useful for the target learner group such as inappropriate language like profanity (for instance, *up/fuck* was removed) or language related to sex (for instance, *oral/sex* was removed) were also scanned for manually and removed. Chronological data dispersion and range dispersion issues were not considered at this stage because of the fact that these two criteria were planned to be examined at a later date. This resulted in a list of 12,615 pairs being included (see Appendix 6). This list was scanned by an experienced, native-speaking teacher of English for general usefulness, and approximately 90 percent were found useful and worthy of direct teaching. Because of the large number of items and the difficulty in recruiting qualified individuals for such a time consuming task, only one judgment was given for the items. Ideally, multiple judges should be used but for this current study this simply was not possible because of a lack of manpower and extremely time consuming task at hand. Despite this, it was confirmed that the

frequency cut-off was not too inclusive or too exclusive, at least to the extent that one experienced, native-speaking teacher of English was concerned.

Discussion

One of the necessary steps to identifying high-frequency collocations/MWUs is to set a frequency cut-off. The frequency cut-off utilized in this study resulted in very good coverage of high-frequency vocabulary, in that 96.75 percent of the tokens of the lemma pairs identified fell within the top 3,000 word families. The lemma pairs also exhibited good coverage of the top 3,000 word families, with 83.14 percent of the word families being represented in the lemma pair list.

However, the large number of items identified presents a challenge. The vast majority of items were deemed useful, even in the lower frequency range of one occurrence per million running words. In fact, this study found that useful collocations can still be found as low as one occurrence per hundred thousand tokens, such as *nice/vacation*, *finish/workout*, and *tend/exaggerate* (Davies, 2010). However setting a more inclusive frequency cut-off would then create a list consisting of more than 2-3,000 word families, which would not be practical in terms of direct instruction. This abundance of useful items poses a serious barrier both research and the study of collocation/MWUs. Therefore, further steps to focus on items with higher learning burdens, or items that have more usefulness for specific learning contexts, must be taken. Such steps include dispersion data analysis, L1 congruency analysis, and semantic transparency analysis.

Conclusion

Determining the extent that frequency data can help inform useful collocation selection revealed that this measure can help inform such selection to an extent, but that there are limitations. First, it was shown that it is possible to set a frequency cut-off that results in a list of collocations that can be practically taught. What at first seemed an impractical amount of items to teach was in reality only 2,789 word families combining with each other in 12,615 different ways, which is within the 2,000 to 3,000 word family estimate of what can be taught directly. And while many useful collocations do occur beyond the frequency cut-off of this study, a list of

collocations resulted that showed very good coverage of high-frequency vocabulary (83.14%) in addition to having 96.74% of the word families within the pairs being within the most frequent 3,000 word families. However, a number of other steps must still be taken to make the data practically usable, despite these positive results.

There was the issue of removing duplicates, or instances when a collocate of one pivot word is also a pivot word. This is a time consuming, manual process that is essential. Moreover proper nouns also need to be removed. This step is also time consuming because it must be done manually. It was also difficult to judge whether a lemmatized collocational pair is part of a larger proper noun without examining concordance data.

RQ2: To what extent is corpus dispersion data useful for identifying MWUs that are deemed worthy of instruction by native English speaker intuition?

Introduction

Because of the large number of collocations that exist, researchers such as Nation (2001) recommend delimiting items selected for direct study to those which have the highest value for learners. If a learner's goal is to master general English, then that learner should only focus on collocations which occur in a balanced way among a variety of genres. Researchers recommend using this as a criterion for study item selection. However, it remains to be seen how corpus data can specifically be used to accomplish this task. Thus, the following experiment will determine whether or not corpus dispersion data is useful for identifying MWUs most representative of high-frequency lemmatized concgrams.

Materials

In this experiment, the data source was the remaining 12,615 lemmatized concgrams (Appendix 6) that were remaining after the completion of research question 1. Dispersion data were sourced from the COCA.

Procedure

Dispersion data for each concgram was collected from the COCA. Its interface allows users to extract dispersion data for five genres: spoken, fiction, magazine, newspaper, and academic. The interface also allows for the extraction of data in 4-year increments: 1990-94, 1995-99, 2000-04, 2005-09, and 2010-12. Since the four-year section 2010-13 was to be completed at the time of this experiment, dispersion data from that section was not included.

A range of parameters for determining balanced distribution were tested due to the gap in research with the corpus used in this study. As with frequency cut-offs, any cut-off set for dispersion or chronological data will also be unavoidably arbitrary. For instance, Hwang and Nation (1995) specifically state that their choice of vocabulary occurring in 10 out of 15 sections of the corpora in their study for balanced dispersion was unavoidably arbitrary. Ackermann and Chen (2013) also arbitrarily just chose inclusion criteria for determining collocations to be the existence in five or more texts in their dataset. Deciding on a parameter that designates a collocation as having balanced/unbalanced dispersion distribution is clearly impossible. Thus, this study experimented with parameters that best approximate balanced distribution.

The parameters utilized required that a specific percentage of the total occurrences had to occur in a majority of the *COCA*'s genres: three or more out of the five genres. First, the intuition of one native English speaker with over ten years' experience teaching Japanese learners was used to determine the best percentage cut-off. Ideally, more than one experienced native speaker should have been used, as Simpson-Vlach and Ellis (2010) did, but this was not possible due to practical limitations of the very time consuming work involved and a lack of volunteers. The lemma list was examined for items specialized in nature, and a number of these items were found to have approximately 5 percent or less of their occurrences in three or more of the genres. Thus, dispersion data was analyzed at three separate percentages to determine the most useful parameter: less than 10 percent, 5 percent, and 2.5 percent of total occurrences in three or more genres. Then pairs flagged at these parameters were examined to determine if they truly were specialized by a native speaker, and thus not worthy of direct instruction for a general English course.

To accomplish this, all flagged items in the list were analyzed to determine if the parameters were not able to identify items that were actually specialized. Ideally multiple examiners should have been used, but the analysis was conducted by myself alone, a native

English speaker with over 10 years' experience teaching English as a second language, again because of the vast number of items and extreme amount of time involved in the analysis. To determine the extent that the dispersion data distribution cut-offs truly identified items that were not worthy of direct instruction, the collocates were judged using my native speaker intuition and teaching experience as guides in regards to their usefulness. Each item was given a rating (see Table 11 below) in regards to its value for learners of general English.

Table 11

System for rating the value of collocates for learners of general English

Rating	Value in regards to direct teaching
1	Provides no value whatsoever if directly taught
2	Provides little value if directly taught
3	Provides questionable value if directly taught
4	Provides value, but with limitations if directly taught
5	Provides clear value if directly taught

After being rated, any items flagged by each of the cut-off parameters that were rated 1 or 2 were tallied. Furthermore, any items not flagged by the cut-off parameters that received ratings of 1 or 2 were also tallied. These two steps would then be used to judge the cut-off parameter's ability to identify collocations that truly are of little or no use for general learners of English in regards to balanced dispersion. Finally, all items identified as being unbalanced that were not flagged were examined to determine if they fell into a common genre (e.g., academic language).

Results

Out of all three parameters tested, the 2.5 percent or more cut-off in three or more genres was shown to be the most useful in regards to both properly flagging items of little use for learners of general English (Figure 1 below), and the 2.5 parameter was also the lowest in

regards to total items either erroneously flagged or judged unbalanced by the native speaker which were not flagged (Figure 2 below). For example, the following items were not flagged by the following parameters but were judged by a native speaker not be worthy of direct instruction for learners of general English: *capital/gain* (10), *charter/school* (5), and *welfare/reform* (2.5). Furthermore, some items were flagged by the three parameters but judged by a native speaker to be worthy of direct instruction. They were: *personality/trait* (10), *look/pale* (5), and *let/ask* (2.5). The most useful parameter was at 2.5 percent, where 347 of the 720 items flagged (48.1 percent) were judged to be accurately flagged by the native speaker (see Appendix 7). The next most useful parameter was at 5 percent, where a total of 538 of the 1,426 items flagged (37.7 percent) were judged to be accurately flagged by the native speaker (see Appendix 8). The parameter that proved the least useful was at 10 percent, where a total of 664 of the 3,193 items flagged (20.8 percent) were judged to be accurately flagged by the native speaker.

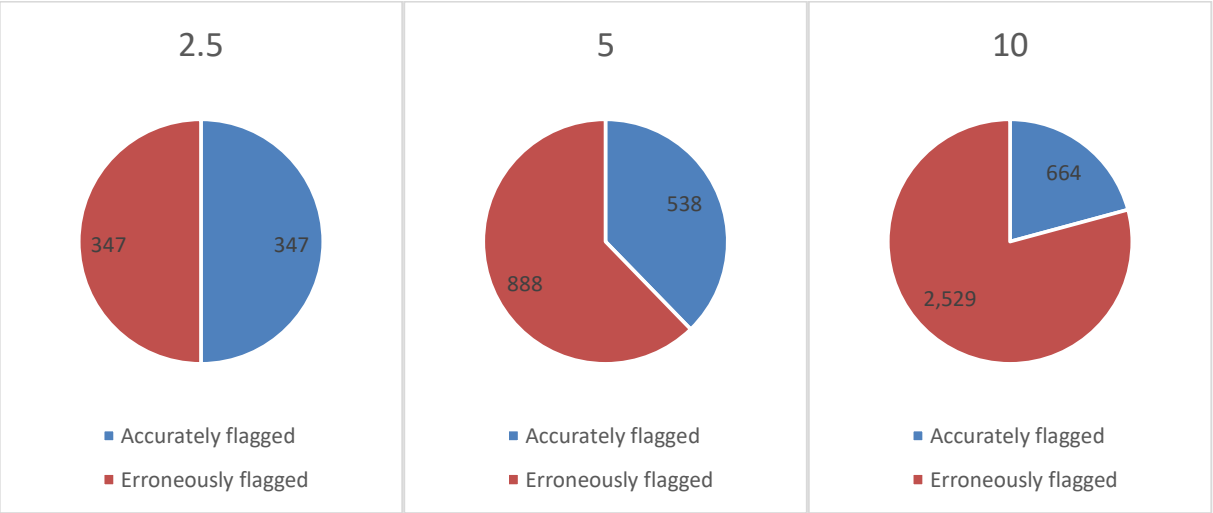

Figure 1. Percentage of items accurately and erroneously flagged for balanced dispersion data distribution at all three parameters

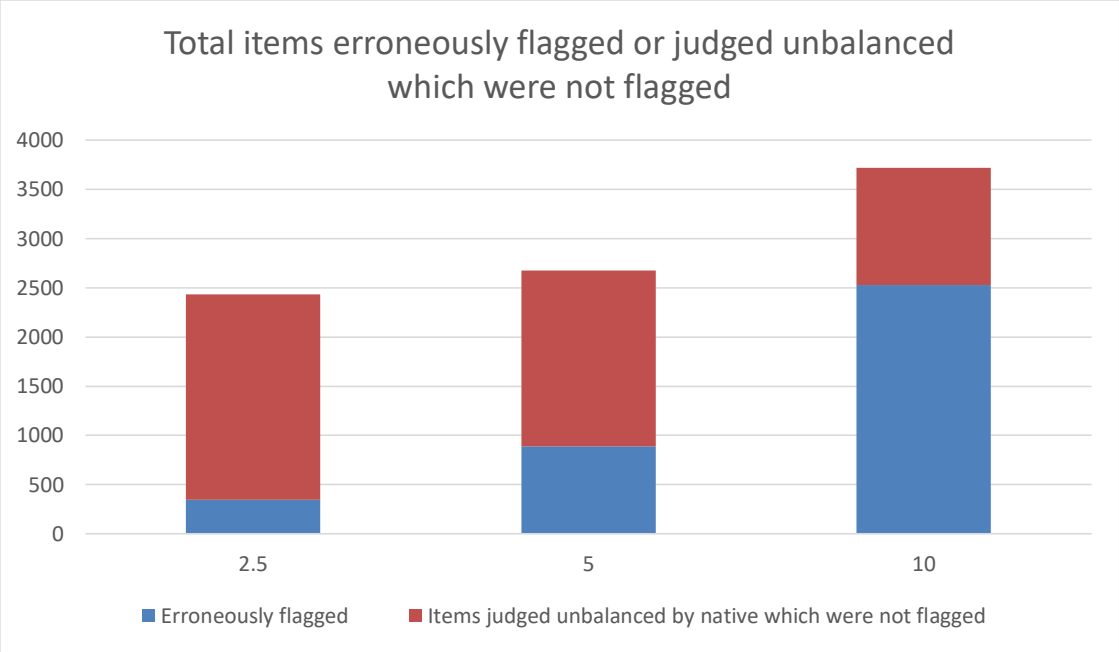

Figure 2. Total items erroneously flagged or judged unbalanced which were not flagged

At the 2.5 percent parameter, 2,088 items were not flagged which were considered to be of low value for learners of general English for dispersion-like reasons (see Appendix 10). At the 5 percent parameter, 1,788 items were not flagged which were considered to be of low value for learners of general English for dispersion-like reasons (see Appendix 11). At the 10 percent parameter, 1,193 items were not flagged which were considered to be of low value for learners of general English for dispersion-like reasons (see Appendix 12).

In addition to the above results, new discoveries were made when the native speaker analyzed the entire list of items. Certain items were deemed to be not worthy of inclusion for learners of general English because they were either inappropriate language (language related to sex, profanity, etc.), grammatical formulations (*so*/*and*), duplicates (how *disease/transmitted* and *disease/sexually* both result in the most common MWU identified being *sexually transmitted disease*), and compound nouns (*log/cabin*, *peanut butter*, etc.) (see Appendix 13). The total amounts can be seen in Table 12 below.

Table 12

Items found to not be worthy of inclusion because they were either inappropriate language, grammatical formulations, duplicates, or compound nouns

Inappropriate	Grammatical	Duplicates	Compound Nouns
15	200	407	129

When items were judged by the native speaker to determine their type of specialized language, four specific types accounted for the vast majority of items: academic language, descriptive language primarily used in fiction, language related to food, and language used primarily on television. Table 13 below gives five samples of the items flagged in each of the four most common types of language at all three parameters.

Table 13

Samples of items flagged for having unbalanced dispersion in each of the four most common genres at all three parameters

Pair	Parameter	Type	Spo.	Fic.	Mag.	News.	Acad.
control/locus	2.5	scientific	0	1	7	1	888
standard/deviation	2.5	scientific	2	4	42	26	2,412
variable/dependent	2.5	scientific	0	0	3	0	2,160
analysis/regression	2.5	scientific	6	1	14	5	1,707
study/longitudinal	2.5	scientific	7	2	64	7	901
slice/thinly	2.5	food	5	5	1,028	402	0
large/skillet	2.5	food	1	6	1,080	347	2
carbohydrate/gram	2.5	food	12	0	567	805	6
flour/cup	2.5	food	0	11	882	484	0

heat/simmer	2.5	food	2	3	833	414	5
lip/lick	2.5	descriptive	13	584	53	9	9
head/jerk	2.5	descriptive	4	597	36	10	4
face/turn	2.5	descriptive	40	1,583	105	44	44
hand/slide	2.5	descriptive	9	644	79	12	11
arm/touch	2.5	descriptive	14	623	52	13	11
moment/commercial	2.5	television	2,785	4	0	0	1
begin/clip	2.5	television	5,874	4	3	8	1
break/welcome	2.5	television	1,250	1	5	6	0
join/studio	2.5	television	829	2	9	16	1
continue/prime-time	2.5	television	510	0	2	1	1
status/socioeconomic	5	scientific	20	8	46	34	998
population/density	5	scientific	16	5	95	29	499
representative/sample	5	scientific	20	5	58	10	499
social/structure	5	scientific	47	38	174	60	1,169
model/predict	5	scientific	30	6	125	33	485
cup/sugar	5	food	40	69	2,836	1,317	2
fat/saturated	5	food	107	9	1,656	2,409	33
heat/medium	5	food	9	4	2,604	983	3
cup/butter	5	food	11	23	1,442	465	0
teaspoon/vanilla	5	food	10	12	1,096	442	0
head/cock	5	descriptive	10	1,057	81	28	7
lip/purse	5	descriptive	5	715	49	15	5
head/tilt	5	descriptive	19	1,311	169	46	26
lip/bite	5	descriptive	29	1,065	53	38	10

mouth/corner	5	descriptive	10	856	72	20	21
commercial/break	5	television	17,903	50	55	45	5
morning/join	5	television	1,770	22	25	30	6
report/correspondent	5	television	712	0	24	47	10
today/guest	5	television	548	17	17	13	0
continue/commercial	5	television	639	4	20	25	23
social/science	10	scientific	80	37	294	277	3,310
waste/solid	10	scientific	26	9	232	138	1,426
social/order	10	scientific	19	56	184	71	1,022
management/water	10	scientific	9	2	59	59	585
soil/erosion	10	scientific	9	3	124	32	396
juice/lemon	10	food	151	80	2,640	1,352	12
high/heat	10	food	64	45	1,461	555	51
oil/large	10	food	117	30	1,141	526	148
acid/fatty	10	food	36	2	912	76	196
large/pot	10	food	7	81	611	358	33
eye/roll	10	descriptive	130	2,389	308	251	42
out/arm	10	descriptive	154	1,899	416	133	75
lay/hand	10	descriptive	89	1,125	181	71	101
head/bow	10	descriptive	62	1,001	120	106	39
hand/clutch	10	descriptive	9	633	53	47	11
cover/story	10	television	812	56	332	121	51
station/public	10	television	438	31	55	169	47
show/tonight	10	television	535	35	30	69	1
columnist/syndicated	10	television	465	7	89	71	16

tape/show	10	television	321	23	50	119	25

A large number of items were judged erroneously flagged by the native speaker. That is, the native speaker felt these items did have value for learners of general English. Table 14 below provides a sample of these items at all three parameters.

Table 14

A sample of pairs flagged for having unbalanced dispersion at all three parameters judged to be erroneously flagged by a native speaker

Pair	Parameter	Section	Spo.	Fic.	Mag.	News.	Acad.
ago/moment	2.5	spoken	1,416	318	46	4	30
good/evening	2.5	spoken	4,592	420	44	51	12
level/significantly	2.5	academic	15	4	65	13	506
indicate/difference	2.5	academic	2	1	23	4	641
effect/significant	2.5	academic	48	6	96	54	2,487
well/obviously	5	spoken	964	72	24	28	19
afternoon/good	5	spoken	882	239	44	26	3
right/absolutely	5	spoken	1,459	134	88	76	25
back/welcome	5	spoken	5,599	185	87	169	21
important/implication	5	academic	20	1	50	7	560
think/definitely	10	spoken	619	40	78	138	11
very/strongly	10	spoken	939	34	89	135	105
question/interesting	10	spoken	630	62	109	64	243
turn/back	10	fiction	846	6,801	924	645	345
high/level	10	academic	128	17	44	47	333

In addition, there was a large number of items judged by the native speaker to be specialized and of little use to general learners that were not flagged at any of the three parameters. Table 15 below provides of sample of such items.

Table 15

A sample of pairs judged to be of little use to general learners not flagged for having unbalanced dispersion by any of the three parameters

Pair	Genre	Spo.	Fic.	Mag.	News.	Acad.
budget/congressional	political	396	1	172	318	111
baseball/bat	sports	177	308	139	153	29
bake/cookie	food	52	68	352	142	7
bond/junk	business	202	4	310	203	9
bone/marrow	medical	368	124	263	312	143

Discussion

Considering a collocational pair's general value in regards to its usefulness across multiple genres proved to be an important criterion; the parameters utilized aided by manual checking identified 1,413 of the 12,615 pairs (11.2 percent) as not being of significant value to general learners of English. However, dispersion data alone was not sufficient in identifying unbalanced items. Often the parameter set either was too inclusive or not inclusive enough, and thus items would be included that were of little value or items of little value were not identified for removal. The most useful parameter was shown to be a cut-off of 2.5 percent of occurrences across three or more genres. While the parameter was useful in helping to flag items to reconsider, native speaker judgments were unavoidable. The parameter could only flag 48.1 percent of the items that were truly of little value.

The largest group that had unbalanced dispersion data was pairs occurring mostly in the academic section. While these pairs would be highly useful for students who plan to do scientific research or read academic journals, such items may not be useful for more general language needs. Thus identifying such genre-specific, unbalanced items can be extremely valuable, either to exclude them or even focus on them if appropriate.

The same can also be said for the large number of pairs that occurred mostly in the fiction section. They consisted of language employed by fiction writers to describe what the reader cannot see. Thus these items do not occur often in any other genres. Again, their inclusion or exclusion depends on the course of study.

Biber, Conrad and Reppen (1998) reminded us that large corpora can skew the type of data we are looking for. This was evident in the disproportionate amount of collocations related to cooking found in the magazine and newspaper sections. Since the magazines and newspapers sourced by the *COCA* regularly featured recipe articles, such items had disproportionate frequency totals. The pedagogical value of directly teaching such items to general learners is questionable except for those who plan to work in the food industry. Thus despite their high frequency, their pedagogical value is in doubt.

Items mostly occurring in the spoken section were also apparently influenced by the data source. The *COCA* sourced much of its spoken section data from television, and in particular, news or talk shows. Thus, the vast majority of the items with unbalanced dispersion in the spoken section consisted of the language newscasters or talk show hosts use, such as commercial break transitions, etc. The value of such items for learners of general English is also arguably low for second language learners, and their discovery shows the importance of dispersion data.

Also of note is how the *COCA* divides its genres, and the effects that has on dispersion data. While much academic and fiction-related language was easily identified, the same cannot be said for other specialized genres, such as business-related collocations, despite it being a clearly specialized genre. Business-related terms were distributed throughout the spoken, magazine, and newspaper genres of the *COCA*, but not in particularly high frequency counts in comparison with academic language, which had its own dedicated genre. Only a small portion of the spoken, magazine, and newspaper genres took its data from business-related sources, such as financial magazines. If the *COCA* were designed with this in mind, such language could have

also been easily identified. Such data would be of clear value to the many learners of business English.

Conclusion

In summary, the data analysis showed that the most useful parameter was able to identify items deemed to be of little value for learners of general English by the native speaker only 48.1 percent of the time. Thus in regard to the extent to which dispersion data can identify what native speakers deem as useful collocations, this experiment revealed that it is limited in that the best parameter was only able to identify about half of the items that needed to be excluded. Since the use of collocations in English materials is more diverse and unpredictable than that of vocabulary, native speakers' judgment is necessitated.

RQ3: To what extent is corpus chronological data useful for identifying MWUs that are deemed worthy of instruction by native English speaker intuition?

Introduction

Because of the large number of collocations that exist, researchers recommend delimiting items selected for direct study to those which have the highest value for learners. If a learner's goal is to master general English, then that learner should only focus on collocations which occur in a balanced way over time. It is obvious that learners should not spend time on items which are dated, too modern, or only occurred during a limited time period. However, research on the usefulness of this criterion in regards to identifying useful collocations has not been conducted to date, and thus questions still remain in regards to its usefulness, and how such a task can be accomplished. Therefore, the following experiment will determine whether or not chronological data from a corpus can be relied upon to help in identifying MWUs most representative of high-frequency lemmatized concgrams.

Materials

In this experiment, the data source was the remaining 12,615 lemmatized concgrams that were remaining after the completion of research question 1. Chronological data was sourced from the COCA.

Procedure

Chronological data for the identified collocates was first collected from the COCA in the same 4-year sections as in research question 2's experiment. First, the intuition of one native English speaker with over ten years' experience teaching Japanese learners was used to determine the best percentage cut-off. Ideally, more than one experienced native speaker should have been used but this was not possible due to practical limitations of the very time consuming work involved and a lack of volunteers. The lemma list was therefore examined using native speaker intuition for pairs which were either dated, too modern, or only occurred during a specific time period. Very few such items existed, but the items that were found had approximately 5 percent or less occurrences in one or more of the four chronological sections. Just as dispersion data was analyzed, chronological data was also analyzed to find items having less than 10 percent, 5 percent, and 2.5 percent of total occurrences in one or more sections. Then pairs flagged at these parameters were examined to determine if they truly were dated, too modern, or not useful because they only occurred during a specific time period by the native speaker, and thus not worthy of direct instruction for a general English course. Next all remaining items in the list were also examined by the native speaker to determine if the parameters were unable to identify items that were dated, too modern, or had little value because they only occurred during a specific time period.

Finally, to determine the extent that the chronological data distribution cut-offs truly identified items that were not worthy of direct instruction, the collocates were then judged by the native speaker in regards to their usefulness just as they were in research question 2's procedure.

Results

Out of the parameters tested, all three were shown to be unreliable from a native speaker's perspective in that approximately 80 percent items of flagged as having unbalanced chronological data dispersion were judged to be erroneously flagged by the native speaker in all

three parameters (see Figure 3 below). At 2.5 percent, only 15 of the 73 items (20.5 percent) flagged were judged to be flagged accurately (see Appendix 14). At 5 percent, only 28 of the 163 items (17.2 percent) flagged were judged to be flagged accurately (see Appendix 15). And at 10 percent, only 67 of the 335 items (20.0 percent) flagged were judged to be flagged accurately (see Appendix 16). Only 5 items beyond the parameters tested were judged by the native speaker to be of little use for learners because of chronological issues (see Appendix 17). Only 67 out of 12,615 items (0.53 percent) were found to have little use for learners because of chronological issues.

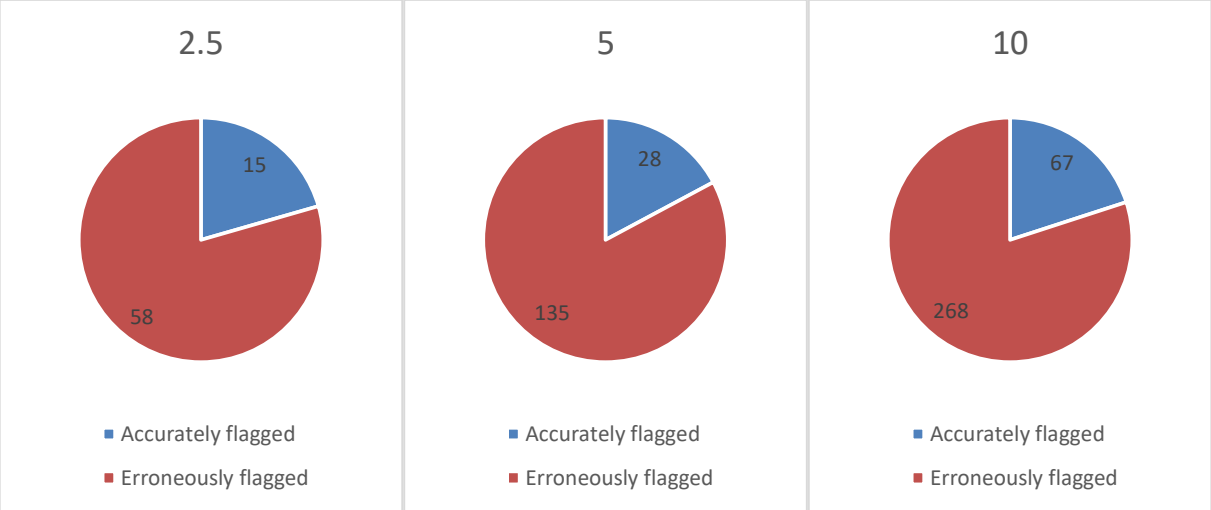

Figure 3. Percentage of items accurately and erroneously flagged for balanced chronological data distribution at all three parameters

In Table 16 below, a sample of some of the items flagged at all three parameters as being of little value to learners of general English because of their chronological data dispersion imbalance, and the 5 items deemed to have chronological issues by the native speaker that were not flagged by any of the parameters are shown.

Table 16

Samples of items accurately flagged at all three parameters (2.5, 5, and 10) and items judged to have chronological issues not flagged by any of the parameters (X)

Pair	Parameter	1990-1994	1995-1999	2000-200	2005-2009
marriage/gay	2.5	10	159	608	527
budget/amendment	2.5	155	445	7	7
suicide/bomber	2.5	13	103	624	615
cell/embryonic	2.5	7	24	375	401
package/stimulus	2.5	251	21	141	554
fund/hedge	5	57	294	268	797
health/reform	5	1050	241	61	939
force/coalition	5	255	27	451	281
new/millennium	5	31	387	422	123
bond/junk	5	535	99	57	32
saving/loan	10	1312	197	95	113
industry/tobacco	10	263	563	194	59
rain/acid	10	427	204	149	83
change/regime	10	63	55	371	238
word/processor	10	282	123	53	74
trade/deficit	X	424	354	121	174
federal/deficit	X	392	109	92	132
federal/insurance	X	345	84	66	139
land/reform	X	156	158	78	290
health/universal	X	169	62	95	279

Discussion

Considering a collocational pair's balanced chronological data distribution, when determining its value for learners, proved to be much less effective than the dispersion data analysis, since only 0.53% of the 12,615 pairs were found to have chronological issues which would make them not worthy of direct study for learners of general English. Furthermore, each parameter was shown to be quite inaccurate in that the vast majority of the items it flagged as having unbalanced distribution was deemed valuable for learners of general English.

Often items erroneously flagged by the parameters were new collocations deemed by the native speaker to have high potential to be used regularly in the future, such as *internet/access.* The types of items that were accurately flagged or deemed by the native speaker to have chronological issues were mostly related to temporal events, such as with *new/millennium*. Items with sudden surges in frequency counts were mostly connected to political events, wars, or such time-sensitive events.

Some items were also deemed too modern, so their future value was unclear. For instance, *cell/embryonic* was flagged by one of the parameters and considered by the native speaker to be of questionable value. It may have high frequency counts simply because it is a new technology and being discussed often, and it is unclear how whether the collocation will continue to be used. The science may become commonplace or outdated, and thus the term may not be discussed as often in the future.

Only a few items were considered as dated, such as *word/processor*. Notably the corpus only provides data back to 1990. If older data were available, then there would be more dated collocations identified. However, within the data's 19-year span, very few dated collocations were found. In addition, if a more detailed chronological breakdown of data were available (i.e., a breakdown by year instead of 4-year sections), a more in-depth analysis would have been possible.

Conclusion

This experiment clearly demonstrated the limited efficacy of chronological data analysis. Not only was there a very small number of items that actually had chronological issues, all of the parameters tested were highly inaccurate, thus again requiring native speaker judgment. Thus this criterion was shown to be of limited value for useful collocation identification.

RQ4: To what extent is consideration for colligation an important criterion for identifying MWUs that are deemed worthy of instruction by native English speaker intuition?

Introduction

Three main approaches (semantic, structural, and lexical) to researching collocations exist, with each approach having virtues and limitations. As discussed earlier, the lexical approach has advantages over the other approaches for this current study. However, if used in a focused way, the ability of the structural approach to consolidate data via grammatical matrices certainly has the potential to improve upon this study's results. How to achieve this and the extent of the improvement are important questions to examine. Therefore, this next experiment determined whether consideration for colligation is an important criterion for identifying MWUs most representative of high-frequency lemmatized concgrams.

Materials

This experiment utilized the list of 11,208 high-frequency lemmatized concgrams remaining after the results of the experiments in research questions 2 and 3 where items which were deemed to have unbalanced dispersion and chronological stability were manually checked for and removed.

Concordance data for each of the 11,208 lemmatized concgrams was collected from the COCA to identify the MWUs most representative of them. This study's approach necessitated the writing of custom concordance software to identify the most common MWUs. Using normal concordance software, such as Anthony's (2011) *AntConc,* was not an option because this study aimed to identify only MWUs in which the lemma occurred, a function not possible with *AntConc* or other concordance software. Furthermore, the large amount of data (over 11,000 pairs) required a batch processing option, another feature not possible with current concordance software. Thus this study used the custom concordance software *AntWordPairs* (Anthony, 2013), a program written specifically for this study. It utilizes Someya's (1998) *E-lemma list*. For coding purposes, Someya's lemma list could not contain duplicate entries, and thus was modified to remove homonyms. For part of speech tagging, the software *GoTagger Version 0.7*

(Goto, 2005) was utilized, and for colligational marker substitution, the software *Textcrawler* (Digital Volcano, 2011) was utilized.

Procedure

The first step was to collect concordance data (example sentences) for each of the 11,208 lemma pairs. Lemmatized concordance searches were conducted, using the COCA's online concordance interface, to identify instances when the collocate occurred either four words to the left or right of the node word in MWUs up to seven words long. The COCA's interface provides options for 100, 200, 500, or 1,000 example sentences to be extracted. Since more data provides more accurate results, this study began by collecting 1,000 example sentences for each pair. However, because of COCA download limits, and the time required for the sentences to load, 1,000 sentences was deemed impractical. However, to ensure that 500 example sentences provided similar data as 1,000 sentences would, results from ten random lemma pairs were compared using both 500 and 1,000 example sentences. Starting with pairs which had frequency counts of 1,000 or more, every 500th pair was selected from the list which was sorted by frequency. Extracting 500 example sentences per lemma pair essentially created a mini corpus for each pair consisting of approximately 13,000 words per pair.

The next step was to identify specific categories of lexical items that occur in high-frequency that could be substituted with colligational markers. Sinclair (1998) defines colligation as the attraction between a lexical item and a particular grammatical category. But as stated earlier in the example of how [*adjective*] + *tea* is useful to an extent but cannot explain why *powerful tea* is not an option while *strong tea* is and thus learners cannot avoid such potential errors, such operational usage of the criterion can become problematic. Furthermore, utilizing a colligational analysis of data with such broad grammatical categories is also quite problematic for a study such as this in that this study aims to pinpoint exact examples learners should study. Thus, this study limited the grammatical categories it considered in its colligational analysis to only those which had the potential to produce results native speakers deemed more useful in identifying MWUs most representative of lemmatized concgrams. For instance, the grammatical category pronouns is a perfect example. When instances of a MWU are counted, such as *buy him a present, buy her a present, buy me a present, buy them a present*,

etc., it is more appropriate to count all pronouns as one 'colligational marker' (as 'object pronoun') to help identify the MWU most representative of how the lemma *buy/present* occur together because all of these examples are essentially the same MWU, albeit simply with different pronouns.

Essentially the goal was to experiment with a number of items that could be substituted with a marker that does not impede the meaning of the MWU as a whole, while providing frequency counts which achieve the goal. However, since no previous research existed, a number of items needed to be chosen and experimented with. A MWU search was conducted on all 11,208 lemma pairs without consideration for collocation. A scan of the full data by a native English speaker revealed that pronouns were one grammatical category that occurred quite often in the MWUs identified, and could easily be substituted without disruption of the meaning of the MWUs as a whole and deemed an improvement upon data analysis from the perspective of native speaker intuition. Thus, steps were taken to substitute the various types of pronouns with markers. In addition, a number of other word categories were also deemed to have similar potential for their results to be improved upon in the colligation treatment: *months, days of the week, ordinal numbers* and *cardinal numbers.*

To use the colligational categories, adjustments for homonyms in the corpus data was necessary. This was done by part of speech tagging using the software *GoTagger* and making replacements using the software *Textcrawler.* First, all instances of the pre-nominal possessive pronoun *her* were changed to *his* as to not interfere with the object pronoun *her.* Then, instances of the ordinal number *second* were changed to 2nd as to not interfere with the noun *second.* Next, instances of the nominal possessive personal pronoun *his* were changed to *hers* to not interfere with the pre-nominal possessive pronoun *his.* Then, the nominal possessive personal pronoun *mine* was replaced with *yours* to not interfere with the noun *mine.* Furthermore, instances of the month *May* and *March* were replaced with *January* to not interfere with the auxiliary verb *may* and the verb *march,* respectively. In addition, the day of the week abbreviations *Sun, Wed,* and *Sat* were replaced with *Mon* to not interfere with the noun *sun* and the verbs *wed* and *sat,* respectively.

Then, *Textcrawler* was used to replace all the pronouns, months, days of the week, ordinal and cardinal numbers with distinct colligational markers in each mini-corpus. The data

was then processed with *AntWordPairs* (Anthony, 2013) to identify the most common MWUs each lemma pair occur in. Because the amount of resulting data was excessive and problematic for the software to process (as explained earlier in this study's *Instruments* section), and the cut-off decided upon by far provided robust enough data to accomplish the study's goals, only MWUs occurring in five percent or more of the corpora were collected. Furthermore, a limit of seven words was set for the length of MWUs that would be chosen to represent each lemma pair.

The next step was a random sample of the MWUs that were affected by the colligational treatment, and a concordance search with the original data not treated for colligational to judge whether a different MWU was identified.

Results

Data from ten random concordance searches was examined for differences between using 500 and 1,000 example sentences. Between the two amounts, the same top MWU was identified for every pair examined, regardless of whether 500 or 1,000 example sentences were used. The data also shows that the frequency counts varied very little when comparisons were made. Table 17 below shows the top MWU identified for each of the ten pairs examined.

Table 17

The top MWU identified when 500 and 1,000 example sentences were utilized

Lemma	POS	Lemma	POS	Multi-word Unit Identified	% out of 500 sentences	% out of 1,000 sentences
announce	verb	week	noun	announced last week	21.6	20.0
trade	noun	deficit	noun	trade deficit	85.6	84.7
body	adj.	upper	adj.	upper body	87.2	86.2
up	adv.	high	adv.	high up	70.0	66.5
little	adv.	better	adv.	little better	100	97.5
court	noun	hold	verb	court held	40.2	42.5
take	verb	charge	noun	take charge	46.4	38.7

care	verb	people	noun	people who care	15.4	10.8
get	verb	look	noun	get a look	23.2	15.7
too	adv.	often	adv.	too often	57.4	33.4

After the initial concordance search, distinct categories of words were found to occur frequently in the MWUs identified. The vast majority of these were pronouns. Thus colligational markers were created for the following types of pronouns:

1. Pre-nominal possessive pronouns (*your, his, her, their, my, our, its*)
2. Subject pronouns (*I, you, he, she, they, we, it*)
3. Object pronouns (*me, us, him, her, them*)
4. Nominal possessive personal pronouns (*theirs, his, hers, yours, mine*)
5. Singular reflexive personal pronouns (*myself, yourself, himself, herself, itself, yourselves, themselves, ourselves*)

It was also determined that four other additional colligational categories should be replaced with colligational markers since they were seen occurring in the original concordance search, did not disrupt the meaning of the MWU as a whole, and could potentially provide more accurate frequency counts. There were:

1. Months (*January, Jan, February, Feb, Mar, April, Apr, May, June, Jun, July, July, August, Aug, September, Sept, October, Oct, November, Nov, December, Dec*)

2. Days of the week (*Sunday, Sun, Monday, Mon, Tuesday, Tue, Wednesday, Wed, Thursday, Thurs, Friday, Fri, Saturday, Sat*)

3. Ordinal numbers (*1st, 2nd, 3rd, 4th, 5th, 6th, 7th, 8th, 9th, 10th, 11th, 12th, 13th ,14th, 15th, 16th, 17th, 18th, 19th, 20th, 21st, 30th, 40th, 50th, 60th, 70th, 80th, 90th, 100th, first, second, third, fourth, fifth, sixth, seventh, eighth, ninth, tenth, eleventh, twelfth, thirteenth, fourteenth, fifteenth, sixteenth,*

seventeenth, eighteenth, nineteenth, twentieth, twenty-first, thirtieth, fortieth, fiftieth, sixtieth, seventieth, eightieth, ninetieth, one-hundredth)

4. Cardinal numbers (*1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38 ,39 ,40 ,41 ,42 ,43 ,44 ,45 ,46 ,47 ,48 ,49 , 50, 51, 52, 53, 54, 55, 56, 57, 58 , 59 , 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1,000, 2,000, 3,000, 4,000, 5,000, 6,000, 7,000, 8,000, 9,000, 10,000, 20,000, 30,000, 40,000, 50,000, 60,000, 70,000, 80,000, 90,000, 100,000, 200,000, 300,000, 400,000, 500,000, 600,000, 700,000, 800,000, 900,000, 1,000,000, one, two, three, four, five, six, seven, eight, nine, ten, eleven, twelve, thirteen, fourteen, fifteen, sixteen, seventeen, eighteen, nineteen, twenty, thirty, forty, fifty, sixty, seventy, eighty, ninety, one-hundred, one-thousand, ten-thousand , one-hundred thousand, one-million*)

It should be noted that these selections are not all-encompassing and other potentially useful colligational patterns may certainly be present in the data. However, due to practical time and computing limitations this thesis could only deal with the above colligational categories and the items listed within them.

After all the mini-corpora were adjusted for homonyms and processed with *AntWordPairs* (Anthony, 2013) to identify the MWUs, and five native speakers who are experienced teachers of English in Japan extracted the MWUs most representative of how each lemma pair co-occurs, the amount of MWUs identified that were affected by the colligational treatment were counted. The results are shown in Table 18 below (see Appendix 18 for a full list of the items).

Table 18

Amount of top MWUs that were affected by each of the colligational treatments

Colligational treatment	Number of top MWUs affected	Percentage of total lemma pairs

Pre-nominal possessive pronouns	259	2.1%
Subject pronouns	208	1.7%
Cardinal numbers	171	1.4%
Object pronouns	74	0.6%
Ordinal numbers	14	0.1%
Singular reflexive personal pronouns	1	0.007%
Nominal possessive personal pronouns	0	0%
Months	0	0%
Days of the week	0	0%
Grand totals	727	5.8%

The colligational treatment for pre-nominal possessive pronouns was shown to be the most common. 2.1 percent of the lemma pairs' top MWUs were affected by this colligational treatment. Treatments for subject pronouns and cardinal numbers also resulted in a significant amount of items being affected. In total, 6.4 percent of all of the top MWUs (719 items) identified were affected by all the colligational treatments conducted. It should be noted that the reason why this total differs from the 727 colligational issues listed in Table 18 above is because eight MWUs had two colligational issues. There are as follows:

1. *put* [pre-nominal possessive pronoun] *in* [subject pronoun] *pocket*
2. *pregnant with* [pre-nominal possessive pronoun] [ordinal number] *child*
3. *released* [pre-nominal possessive pronoun] [ordinal number] *album*
4. *gave* [object pronoun] [pre-nominal possessive pronoun] *card*
5. *celebrated* [pre-nominal possessive pronoun] [ordinal number] *birthday*
6. *celebrates* [pre-nominal possessive pronoun] [ordinal number] *anniversary*
7. *birth of* [pre-nominal possessive pronoun] [ordinal number] *child*
8. *give* [subject pronoun] [cardinal number] *dollars*

Ten random samples were then taken from the top three types of colligation treatment found to affect the top MWU identification. These were then compared to a top MWU search with untreated data. Out of the 30 items selected, only 13 (43.3 percent) resulted in different MWUs being identified. For items affected by the pre-nominal possessive pronoun treatment, only four out of ten top MWUs differed. With the subject pronoun treatment, only three out of ten top MWUs differed. With the cardinal number treatment, six out of ten of the top MWUs differed. These results are summarized in Table 19, 20, and 21 below.

Table 19

Comparison between ten random samples of top MWUs affected by the colligational treatment for pre-nominal possessive pronouns and the results that would have occurred without the treatment. (Note: Items in bold indicate those that showed differences in the top MWU identified, and instances of a slot in which a pre-nominal possessive pronoun exists are represented with "")*

Lemmatized concgram pair	MWU identified with this study's colligational treatment	MWU that would have been identified without this study's colligational treatment
***hand* (noun) *wave* (verb)**	***waved * hand***	***waved a hand***
***live* (verb) *life* (noun)**	***live * life***	***live life***
***base* (verb) *experience* (noun)**	***based on * experience***	***based on experience***
***attention* (noun) *focus* (verb)**	***focus * attention***	***focus attention***
head (noun) *gun* (noun)	*gun to * head*	*gun to his head*
hand (noun) *extend* (verb)	*extended * hand*	*extended his hand*
eye (noun) *wipe* (verb)	*wiped * eye*	*wiped her eye*
life (noun) *ruin* (verb)	*ruin * life*	*ruin your life*
put (verb) *hand* (noun)	*put * hand*	*put her hand*

sit (verb) *desk* (noun)	*sitting at * desk*	*sitting at his desk*

Table 20

Comparison between ten samples of top MWUs affected by the colligational treatment for subject pronouns and the results that would have occurred without the treatment. (Note: Items in bold indicate those that showed differences in the top MWU identified, and instances of a slot in which a subject pronoun exists are represented with "")*

Lemmatized concgram pair	MWU identified with this study's colligational treatment	MWU that would have been identified without this study's colligational treatment
***see* (verb) *mirror* (noun)**	***mirror * saw***	***mirror and saw***
***wear* (verb) *dress* (noun)**	***dress * wore***	***wearing a dress***
***take* (verb) *back* (adverb)**	***take it back***	***take back***
how (adverb) *interact* (verb)	*how * interact*	*how they interact*
get (verb) *when* (adverb)	*when * got*	*when I got*
make (verb) *hard* (adverb)	*makes * hard*	*makes it hard*
could (verb) *suppose* (verb)	*suppose * could*	*suppose you could*
belong (verb) *where* (adverb)	*where * belong*	*where I belong*
think (verb) *pretty* (adverb)	*think * is pretty*	*think she is pretty*
want (verb) *whenever* (adverb)	*whenever * want*	*whenever you want*

Table 21

Comparison between ten random samples of top MWUs affected by the colligational treatment for cardinal numbers and the results that would have occurred without the treatment. (Note:

Items in bold indicate those that showed differences in the top MWU identified, and instances of a slot in which a cardinal number exists are represented with "")*

Lemmatized concgram pair	MWU identified w/ colligational treatment	MWU identified w/o colligational treatment
***get* (verb) *second* (noun)**	***got * seconds***	***seconds to get***
***nearly* (adverb) *decade* (noun)**	***nearly * decades***	***nearly a decade***
***just* (adverb) *year* (noun)**	***just * years***	***just a few years***
***live* (verb) *mile* (noun)**	***live * miles***	***live within 50 miles***
***nearly* (adverb) *mile* (noun)**	***nearly * miles***	***nearly a mile***
***minute* (noun) *second* (noun)**	***minutes * seconds***	***seconds to one minute***
estimate (verb) *percent* (noun)	*estimates that * percent*	*estimates that 80 percent*
divide (verb) *group* (noun)	*divided into * groups*	*divided into two groups*
over (adverb) *month* (noun)	*over * months*	*over six months*
roughly (adverb) *percent* (noun)	*roughly * percent*	*roughly 10 percent*

Discussion

Regarding the amount of data collected to create each mini-corpus used in this study, 500 example sentences were deemed to be able to produce similar results as 1,000 example sentences would when concordance data was compared. The example shown in Table 15 earlier demonstrates that collection of 500 versus 1,000 example sentences for each lemma pair made no difference in identifying the most common MWU. However, collecting the data was a manual process of copy and pasting from the COCA's interface, something it was not designed for. Thus through the process unnecessary data was also copied, and therefore a multi-step process of pasting into an Excel file, then copying only the sentences and pasting again into a Word file, and then saving the file, was necessary to remove this data. Being a cumbersome,

time-consuming process, corpus computer interface and vocabulary learning software creators may want to consider this for future design.

When the initial concordance data was examined after processing the compiled mini-corpora, various types of pronouns occurred quite often within the MWUs identified. Other categories of words, such as cardinal numbers, also frequently occurred. Thus such word categories became the focus of this study's colligation experiment. However, because of a lack of previous research, other categories were experimented with as well. While it is true that not all of these proved fruitful, the resulting data did provide insight as to specific types of colligation that, when addressed, can improve upon the reliability of MWU identification.

The colligational treatment for pre-nominal possessive pronouns was shown to be the most useful. Treatments for singular reflexive personal pronouns, nominal possessive personal pronouns, months, and days of the week did not prove useful; only one item was affected in the entire list by all of these treatments. At first glance, the colligational treatment was shown to be an important step in the identification of the most frequent MWUs, most representative of lemmatized concgrams, in that 719 (6.4 percent) of the total concgrams examined had their most common MWU change. However, when a sample of the MWUs was compared to the MWUs that would have been identified without a treatment for colligation, only 43.3 percent of the items actually had differing results. Therefore, while frequent counts were always improved upon, the treatments did not always end with improved results.

Yet before the colligational treatment could be conducted, homonym interference in the data had to be dealt with. The process was complex, cumbersome, and very time-consuming due to the lack of dedicated software to conduct such a task. It would be useful if software developers considered such functionality and ways to improve the efficacy of conducting such data modification.

Conclusion

This experiment compared results from different sized lemmatized concgram corpora and provided evidence as to the type of results one can expect when conducting specific colligational treatments on data. It showed how 500 example sentences that contain a target pivot word and collocate would produce similar results as 1,000 example sentences would. It also showed that

MWU searches for 6.4 percent of the lemma pairs examined were affected by the colligational treatment taken in this study. However, when a sample of these items was examined more deeply, it was found that nearly half showed no difference in the top MWU identified. For example, the colligational treatment may identify *drive* [object pronoun] *home,* while even without the colligational treatment *drive him home* was identified. These are essentially the same MWU. So, in reality only approximately three percent of the 6.4 percent of items that exhibited colligational issues had their results improved upon. Since the steps needed to achieve these improvements were found to be extremely time consuming and complex, it is clear that there is a need for a more efficient methodology for such colligation treatments. Software designers should thus consider ways to automate some of the steps taken in this study.

This experiment did have its limitations. Due to the lack of previous research and no standard on how to conduct such a data analysis, choices for the types of colligation examined were subjective. Quite possibly other types of colligation exist in the data that could also prove fruitful if treated. Thus more research is needed in regards to other types of colligation that may improve results if treated. Despite these limitations, this experiment did provide new insights into a previously unexplored area of linguistic analysis that certainly has the potential for creating improved resources that help learners achieve fluency in a second language.

RQ5: What percentage of MWUs is deemed by experienced native speakers who teach English as a second language at university worthy of expanding beyond their most frequent exemplar to provide learners with useful information about how the items commonly occur formulaically?

Introduction

Corpora can no doubt help improve upon our ability to select useful language to teach to second language learners. However, current technology does not enable researchers to use corpora alone for the identification of MWUs most representative of a lemma pair. While corpora and concordance software can identifying MWUs and sort them by frequency, they cannot help identify MWUs which would benefit by being expanded beyond their cores. For example, corpora and concordance software can easily identify *come to terms* as the most common MWU of the lemma pair *come/term*, but cannot make a decision regarding whether or

not to expand such an example to include other words that frequently occur beyond this core string, such as with *to come to terms with.* In fact, this was the case in this current study. The corpus data did identify *come to terms* as the top MWU, but native speaker analysis of the data led to the extension of this top MWU to a MWU identified slightly lower in the list: *to come to terms with.* Thus, *to come to terms with* was chosen to be the MWU most representative of the lemma *come/term.* The extent to which this is an issue or not has not been examined in previous research, and thus this important question remains unanswered. Therefore, the next experiment will be conducted to determine the percentage of MWUs deemed by native speakers worthy of expanding beyond their most frequent exemplar to provide learners with useful information about how the items commonly occur formulaically.

Materials

This experiment utilized the same list of 11,208 high-frequency lemmatized concgrams and concordance data as in research question 4.

Procedure

All concordance data collected was processed using *AntWordPairs* (Anthony, 2013), and the data were broken up and distributed among five native English speaking university English language professors in Japan who then identified the most frequent MWU in which the lemma occur in. Then, these native speakers examined subsequent MWUs (sorted by frequency) which also contained this top MWU along with other words to its left or right to determine if extending the MWU to the left or right of this core MWU would provide useful information for learners. Ideally, each item in the list would have been rated by each native speaker but due to the fact that there were over 11,000 items and this step is extremely time consuming, this simply was not possible so there are clear limitations to how the findings can be interpreted.

Native speakers were instructed to use their intuition, knowledge of the English language, and experience teaching English to determine whether or not it was appropriate to extend beyond the core MWU. For instance, if the top MWU identified is *come to terms* and the second *come to terms with*, a native speaker would use his/her practical knowledge to opt to choose to extend and add *with* because of its high frequency of following *come to terms*, the low frequency of any

other options, and the general usefulness of the phrase. Furthermore, native speakers were instructed to also utilize the frequency data available as well. For example, if the top MWU identified was *come to terms* and had a frequency count of 500, and the second was *come to terms with* and had a frequency count of 499, for all practical purposes it is clear that in such a situation it would be best to opt to have the MWU *come to terms with* represent the lemma pair *come/terms.*

Then a random sample of 100 of the final MWUs identified was examined to determine which percentage native speakers extended beyond the top MWU.

Results

Native speakers opted to extend MWUs beyond the core pivot and collocate in 53 percent of the 100 random MWUs sampled (see Appendix 19). For instance, the most frequent MWU for the lemma pair *come* and *term* was found to be *come to terms*, at 243 occurrences (see Table 22 below). However, the next most common string in the data beyond *come to terms* was *come to terms with* (229 occurrences), and beyond that, *to come to terms with* (129 occurrences). Thus a native speaker judged *to* **come to terms** *with* as being the MWU most representative of the lemma pair *come* and *term.* To accomplish this, in addition to utilizing available frequency data, native speakers relied on their intuition to not only add strings that truly represented common usage, but that also provided learners with useful information.

Table 22

MWUs identified from 500 example sentences in which the lemma pair 'come' and 'term' both occur

MWU	Occurrences in 500 sentences
come to terms	243
come to terms *with*	229
to **come to terms**	133
to **come to terms** *with*	129

coming to terms	96
coming to terms with the	86
to **come to terms** *with the*	44
come to terms *with* [pre-nominal possessive pronoun]	28
coming to terms with the	26

Discussion

In regards to the value of extending MWUs beyond the core pivot and collocate, the data suggests that this is an important criterion to consider when attempting to identify MWUs most representative of lemmatized concgrams. Native speakers opted to extend MWUs in more than half of the items examined. Corpus data and software alone cannot accurately identify such extensions, and thus this aspect of the study showed the extent to which data becomes modified when native speaker intuition is referred to for intervention in MWU identification.

Conclusion

This experiment highlighted the value of extending MWUs beyond the core pivot and collocate. Over half of the sample examined were deemed to be worthy of extended beyond the most frequent MWU by native speakers. However, because this is a procedure that cannot be accomplished using software, it can be very time-consuming and thus more research needs to be done to help possibly automate this process somehow.

RQ6: To what extent is semantic transparency an important criterion to consider when attempting to identify collocations deemed worthy of direct instruction by native speakers?

Introduction

Despite there being agreement in regards to the value of collocations, even today there is still much disagreement as to what should and shouldn't be considered to be a collocation. Some researchers believe that words which frequently co-occur but that are also semantically opaque should only be considered collocations (Moon, 1994; 1997; Van der Meer, 1998). This is logical

in that it would help delimit items to only those with a higher learning burden. However, it remains to be seen the extent to which high-frequency collocations are semantically opaque or transparent. Thus, the following experiment will show where the high-frequency collocations identified in this study fall on the spectrum between literal and idiomatic to enable practitioners to know which particular items need additional study time because of the additional learning burden that is added as a collocation moves closer down the spectrum to the idiomatic end.

Materials

This experiment utilized the same list of 11,208 high-frequency lemmatized concgrams that was identified in research question 4.

Procedure

In this study, the list of MWUs was double read by two native English speaking English language teachers to determine their level of semantic transparency. Determining a collocation's level of semantic transparency is not a simple task, and it is essential to recognize that there is a cline of fixity (Kellmer, 1994; Shin, 2006). Grant and Bauer (2004) suggest distinguishing such items along this cline by breaking them down into the following four categories:

1. Literals: A MWU is a 'literal' if the meaning of each word alone is the same as it is when it is paired as a collocation. (e.g., *eat breakfast*)
2. ONCEs (One Non-Compositional Element): If only one of the core words in the MWU is figurative, then that collocation is considered to be a 'ONCE'. (*driven to quit*)
3. Figuratives: A MWU is a 'figurative' when it is not literal, but it is possible to understand the collocation by pragmatically reinterpreting it. (e.g., *hit the nail on the head*)
4. Core idioms: If the whole MWU is figurative, and it is not possible to reinterpret its meaning to understand it, then it is considered to be a 'core idiom'. (*pull someone's leg*)

However, while analyzing the data the raters began to notice items which do not seem to fit within the above categories. Thus, a new category was created:

5. Outliers: When collocations contained a homonym that could easily be misunderstood (when the significantly rarer homonym is used), the collocation was marked as an 'outlier' (e.g., *bear children*). Collocations were also given this rating when they had very specific meanings which learners have a high probability of misunderstanding (e.g., *boot camp, social security, foster care*). In addition, if a collocation seemed to be formed arbitrarily (there is no rhyme or reason why a particular word is used, and not another logical alternative), it was also given this rating.

Examples of outliers include *take measures, deliver a speech,* and *to stand trial.* For instance, why do we *take* measures and not say *create measures*? Why do we *deliver a speech* but don't *deliver gossip*? Furthermore, wouldn't it be more logical to just say *have a trial*? Recognizing these arbitrary ways in which language combines is essential to recognizing learning burden.

After the two raters analyzed all the data and gave each collocation a rating, inter-rater reliability was determined using the percent agreement measure.

Results

Inter-rater reliability was confirmed as only 245 collocations in total were found to have disagreement between the two raters (see Appendix 20). Such items were simply difficult to rate, and could be viewed from different perspectives easily. For instance, *to go to the bathroom* was rated to be a 'literal' by one rater, and as an 'outlier' by the other rater. On one hand, a person can literally be going to the bathroom (the location) itself, but it can also be viewed from the perspective of meaning that a person needs to urinate. The reviewer that rated it as 'outlier' viewed it from this perspective, but in the end it was decided that since this MWU is used in the literal sense the majority of the time it should be rated as 'literal'. As was mentioned in the procedure section of this experiment and also noted by previous researchers, rating semantic transparency is not a simple task by any means, and items such as the example above can end up being difficult to rate. Despite this, at 97.9 percent, the two raters clearly could be relied upon to rate the items in a similar fashion. Any items that there was disagreement on were re-examined and their ratings were adjusted. Table 23 below is a summary of the final results (see Appendix 21 for the full list).

Table 23

Sematic transparency ratings of the collocations (percentage of total items in italics)

Literal	ONCE	Figurative	Core Idiom	Outlier
9,641/*86.0*	676/*6.0*	193/*1.7*	179/*1.6*	519/*4.7*

Discussion

The results of this study revealed that speakers will native-like ability in English considered the vast majority of high-frequency collocations examined (86.0 percent of them) to be literal formulations. As the value of high-frequency items is well-known and that other factors may influence the learning burden of these items (L1 congruency), suggesting that such a large chunk of the language not be taught directly to students as Moon (1994) suggests seems imprudent.

High-frequency vocabulary is ubiquitous. It can cover up to 80 percent or more of the running words in most texts (Nation, 2008). Thus, Nation (2001b) believes such vocabulary deserve direct teaching time. However, how should such vocabulary be taught to learners? In fact, learning collocations rather than isolated words has been found to actually be easier (Ellis 2001). For example, Bogaards (2001) found that multi-word expressions containing familiar words were retained 10% more than completely new single words immediately after a learning session and also 12.1% more in a delayed posttest three weeks later. Therefore, the teaching of high-frequency vocabulary with their common collocates in the form of multi-word expressions that the collocates typically occur within would be ideal. However, such items would be excluded from what is to be taught directly if Moon's (1994) position is followed. Thus, if learners want to study high-frequency vocabulary in the most efficient way possible, semantically transparent collocations must be taught due to the fact that they make up the vast majority of how high-frequency vocabulary co-occurs.

It is true that the learning burden of a literal collocation is low and that semantically opaque collocations deserve more focus in comparison to semantically transparent items.

However, this study provides evidence which shows how using a measure such as semantic transparency alone to select collocates to teach directly can be problematic. Furthermore, in addition to the factor of L1 congruency, this study also highlighted how certain items fall into particular categories that were not utilized in previous researchers' experiments (e.g., collocations which contain homonyms, arbitrarily formed collocations), and thus should be considered in future research with a similar aim. Consequently, using the simple measure of semantic transparency alone may not be useful in that it excludes a large number of collocations which otherwise may deserve direct teaching time.

Conclusion

This study reveals that the vast majority of the high-frequency collocations examined are considered to be literal formulations. This makes using semantic transparency alone as the measure by which teachers identify and subsequently select collocations to teach to students directly problematic because by doing that, much of high-frequency vocabulary thus ends up being excluded from a collocation/multi-word expression-based approach to vocabulary instruction.

This study highlights the danger of utilizing rigid definitions of linguistic phenomenon when grappling with the practical goal of selecting items to teach second language learners. It also reveals some potential new categories that researchers should consider when rating the semantic transparency of a collocation. With this knowledge, teachers and future researchers may be able to improve upon the choices they make in regards to the explicit teaching of high-frequency collocations.

RQ7: To what extent is L1-L2 congruency an important criterion to consider when attempting to identify English MWUs deemed worthy of direct instruction by native speakers to Japanese learners?

Introduction

Researchers agree that L1-L2 congruency is an important factor that affects a word or MWU's learning burden (Gitsaki, 1996; Laufer & Eliasson, 1993; Nesselhauf, 2005). By

identifying such items, researchers can pinpoint specific items which deserve more teaching time. However, to date, there is still a lack of research in regards to the extent to which congruency exists between certain L1s and high-frequency English worthy of direct instruction. Thus, to fill this gap in the research, the following experiment determined the L1-L2 congruency of high-frequency MWUs between English and Japanese.

Materials

This experiment utilized the same list of 11,208 high-frequency lemmatized concgrams that was identified in research question 4.

Procedure

A translator with native-like ability in both the L2 (English) and the L1 (Japanese) gave L1-L2 congruency ratings to each MWU in the list. The rating was from 1-12 points, with 12 points equating to total congruency. A 12 point system was used because the vast majority of MWUs in the list consisted of either three or four words, and thus it was easy to divide this number by three or four. First, the translator counted the number of words in the MWU and divided that by 12. For instance, each word in the MWU *wake up late* would thus be worth 4 points. Then, each word in the MWU was compared to each word in its translation. If a word's literal meaning differed, or it was simply not present in the translation, it was not awarded points. If a word was in the same word family but was a different part of speech, had slight semantic difference, or a combination of both that word was given half its allotted points. If the translation contains an extra word that was not present in the English, then points allotted for one word were subtracted. If one of the English words in the MWU did not exist in Japanese, such as English articles, the translator was instructed to ignore it because this study aimed to only identify the extent to which the L1 has the potential to influence a learner to make an error, and not to judge whether an item had the potential to judge a learner's proficiency in the L2 itself. The translator was also instructed to ignore when there was a different word order because of the different grammar across the languages in question.

Ideally, multiple translators should have been used in this study and inter-rater reliability could be used to validate the results. However, because the task at hand was so time consuming

and the rater needed very high fluency in both languages to able to make accurate judgments, only one such translator could be found to volunteer their time.

Results

The results of the comparison between 11,208 English MWUs and their Japanese translations can be seen in Table 24 below (see Appendix 22 for the full list). 56.4 percent (6,320 MWUs) received a rating of 0-9. 4,888 of the MWUs received a rating of 10-12, with the vast majority of those items (84.8 percent) being considered 100 percent congruent. Thus, approximately half of the items examined were considered from somewhat to totally incongruent with Japanese.

Table 24

L1-L2 congruency ratings of high-frequency English MWUs with Japanese translations (percentage of total items in italics)

Rating	0-3	4-6	7-9	10-12 (12)
	996 *(11.3)*	2,419 *(4.6)*	2,905 *(3.9)*	4,888 *(2.3)* (4,146 *(2.7)*)

Discussion

The results of this study made it salient that a large proportion of the MWUs examined were not congruent with Japanese to some extent. More than half of the items examined pose a higher learning burden because of incongruency with their Japanese translations. Such a large number of items make it clear that L1-L2 congruency is an important factor to consider when choosing English items for Japanese learners to focus on.

As discussed earlier, L1-L2 congruency is clearly an issue for any study which relies on semantic transparency as the sole criterion upon which to judge learning burden when selecting collocations for students to focus on. Such data can be utilized to improve upon the efficacy of learning by, for example, limiting this study's list of 11,208 MWUs to those which are

incongruent to an extent with their translations. This would create a list of items which need additional study time because of the higher learning burden of such items. In addition, such a list could also be useful for learners who have a limited amount of time to study but who want to focus only on items they have a higher chance of making an error with. For instance, if a cut-off of 6 out of 12 points of this study's L1-L2 congruency rating is utilized, the list can be made significantly shorter. The 11,208 items becomes a list of only 3,414 items which half of the MWU differs with its translation (see Appendix 23). Such a reduction in volume could be significant in helping learners achieve fluency in a more efficient manner.

However, this study clearly has limitations to the implications of its findings. Mainly, L1-L2 translation and comparison is not an exact science. There are not only various ways and levels of quality of translation, but there is also an aspect of subjectivity in making L1-L2 congruency ratings. Furthermore, a procedure for conducting such a comparison has yet to be solidified in previous research and thus this study had to create its own rubric with which to judge congruency. Moreover, this study only relied upon one rater and results could have been different if multiple raters were used and inter-rater reliability was conducted. However, as stated earlier, this was not possible due to the difficulty of finding volunteers that were qualified enough to make such language judgments and were also able to handle the extremely large amount of time consuming work. While this study acknowledges these limitations, it should also be made clear that such issues are unavoidable due to the task at hand. Thus, this study best approximates to the fully extent possible the answer to the research question it set forth.

Conclusion

This study made it salient the extent to which L1-L2 congruency affects the learning burden of high-frequency English MWUs. More than half of the 11,208 English MWUs examined were found to be incongruent to an extent with their Japanese translations. Thus, the learning burden of a large proportion of the items examined clearly has the potential to be affected. This large percentage warrants the use of L1-L2 congruency as a criterion in selecting particular items to spend additional study time on to help learners avoid making production errors influenced by their L1. While there are limitations to interpreting the results of this study, such as the lack of multiple raters, it should still be considered as a good step forward towards

improving upon the efficacy of language learning. At a minimum, this study constitutes a first step towards the ultimate goal set forth in this dissertation as a whole, and hopefully more research will be done in the future to corroborate these findings.

RQ8: To what extent is native speaker intuition useful in regards to high-frequency vocabulary usage in context creation?

Introduction

Corpora, by their very nature, are not perfect. For some tasks, it may actually be preferable to rely on an experienced ESL practitioner who is also a native speaker (as was in parts of this thesis). For instance, such an individual may be better suited for the job in comparison to utilizing corpus data analysis when the task is to create example sentences to help teach MWUs because the native speaker can take into account word frequencies in comparison with the target MWUs. This is key to helping students to learn how a word or phrase is used in proper context while not increasing the learning burden of the item.

However, to date, no previous research has examined the extent to which a native speaker's intuition can be relied upon to create example sentences whose contents mostly fall into the high-frequency realm on a large scale. Thus, this next experiment will examine the type of data native speakers create with the simple instructions to write example sentences for high-frequency MWUs using high-frequency supporting context as much as possible while still producing natural, appropriate examples. It was designed to determine whether or not native speakers could be relied upon using only their intuition to accomplish such a task.

Materials

This experiment utilized the same list of 11,208 high-frequency lemmatized concgrams that was identified in research question 4.

Procedure

The 11,208 lemma pairs were distributed among four native speakers—two Americans and two Canadians—who wrote an example sentence for each set of MWUs they were assigned

to. Thus, each volunteer wrote 2,802 sentences. These native speakers were experienced ESL practitioners, each with ten years or more experience teaching English as a second language. Each native speaker was instructed to create an example sentence for each MWU using as much high-frequency vocabulary as possible while still creating natural and appropriate sentences. Essentially, the goal of the native speaker was to create an example sentence that did not increase learning burden, but rather lowered the burden while also highlighting an item's typical usage in the language.

Then, the formulaic sequences alone were processed with Heatley, Nation and Coxhead's (2002) *RANGE* program to determine the extent to which the contents fell into the high-frequency realm. This program combines the BNC and COCA corpora to produce a frequency list in which other texts can be compared to. This frequency list consists of the top 25,000 word families in the combined corpora, along with levels for noise in the data (26-30, 32, and 34) such as non-words *hmm, eh, arrgh,* and random abbreviations such as *AAL*, proper nouns (31), and compound nouns (33). After that, the same analysis was repeated, but with the formulaic sequences within the example sentences created by the native speakers. The results were compared to each other. Finally, the formulaic sequences within the example sentences were processed with Cobb's (2013) *Vocabprofiler* to specifically determine which of the top 3,000 word families were not covered by the data.

Results

Example sentences written by all four native speakers were combined, which in total consisted of 159,211 tokens (see Appendix 24 for the full list). The formulaic sequences alone and the formulaic sequences with the example sentences were examined using *RANGE*, and Tables 25 and 26 below show their coverage of the top 34 groups of 1,000 word families of English.

Table 25

Word family frequency breakdown of formulaic phrases using RANGE

Word Family Frequency	Total Tokens / %	Total Types / %	Families

Level			
1	25,081/78.04	1,942/44.28	923
2	4,445/13.83	1,202/27.41	721
3	2,071/ 6.44	811/18.49	589
4	277/ 0.86	215/ 4.90	202
5	95/ 0.30	84/ 1.92	84
6	38/ 0.12	33/ 0.75	31
7	10/ 0.03	10/ 0.23	10
8	11/ 0.03	11/ 0.25	10
9	4/ 0.01	4/ 0.09	4
10	0/ 0.00	0/ 0.00	0
11	3/ 0.01	3/ 0.07	3
12	2/ 0.01	2/ 0.05	2
13	1/ 0.00	1/ 0.02	1
14	1/ 0.00	1/ 0.02	1
15	0/ 0.00	0/ 0.00	0
16	0/ 0.00	0/ 0.00	0
17	0/ 0.00	0/ 0.00	0
18	0/ 0.00	0/ 0.00	0
19	0/ 0.00	0/ 0.00	0
20	0/ 0.00	0/ 0.00	0
21	0/ 0.00	0/ 0.00	0
22	0/ 0.00	0/ 0.00	0
23	0/ 0.00	0/ 0.00	0
24	0/ 0.00	0/ 0.00	0
25	0/ 0.00	0/ 0.00	0
26	0/ 0.00	0/ 0.00	0
27	0/ 0.00	0/ 0.00	0
28	0/ 0.00	0/ 0.00	0

29	0/ 0.00	0/ 0.00	0
30	0/ 0.00	0/ 0.00	0
31	6/ 0.02	3/ 0.07	3
32	2/ 0.01	2/ 0.05	2
33	60/ 0.19	38/ 0.87	37
34	0/ 0.00	0/ 0.00	0
Not in the lists	32/ 0.10	24/ 0.55	
Totals	32,139	4,386	2,623

Table 26

Word family frequency breakdown of formulaic phrases within example sentences created using native speaker intuition using RANGE

Word Family Frequency Level	Total Tokens / %	Total Types / %	Families
1	136,707/85.87	2,659/33.92	985
2	13,074/ 8.21	1,959/24.99	900
3	5,271/ 3.31	1,357/17.31	785
4	1,120/ 0.70	557/ 7.10	449
5	663/ 0.42	281/ 3.58	248
6	234/ 0.15	143/ 1.82	127
7	101/ 0.06	73/ 0.93	67
8	90/ 0.06	51/ 0.65	48
9	44/ 0.03	33/ 0.42	33
10	35/ 0.02	26/ 0.33	25

11	26/ 0.02	14/ 0.18	12
12	18/ 0.01	9/ 0.11	8
13	6/ 0.00	5/ 0.06	4
14	6/ 0.00	5/ 0.06	5
15	1/ 0.00	1/ 0.01	1
16	1/ 0.00	1/ 0.01	1
17	1/ 0.00	1/ 0.01	1
18	2/ 0.00	2/ 0.03	2
19	0/ 0.00	0/ 0.00	0
20	0/ 0.00	0/ 0.00	0
21	0/ 0.00	0/ 0.00	0
22	0/ 0.00	0/ 0.00	0
23	0/ 0.00	0/ 0.00	0
24	0/ 0.00	0/ 0.00	0
25	0/ 0.00	0/ 0.00	0
26	0/ 0.00	0/ 0.00	0
27	0/ 0.00	0/ 0.00	0
28	0/ 0.00	0/ 0.00	0
29	0/ 0.00	0/ 0.00	0
30	0/ 0.00	0/ 0.00	0
31	753/ 0.47	251/ 3.20	229
32	54/ 0.03	11/ 0.14	8
33	733/ 0.46	221/ 2.82	189
34	36/ 0.02	14/ 0.18	13
Not on the lists	235/ 0.15	166/ 2.12	
Totals	159,211	7,840	4,140

Tables 25 and 26 show that the phrases themselves consisted of 2,623 word families and after the example sentences were written, there were only 1,517 word families added by the example sentences.

Table 27 below shows the percentage of items in the top 3,000 word families of English that were not covered by any of the words in the example sentences.

Table 27

Vocabprofiler breakdown of top 3,000 word family words not covered by example sentences created using native speaker intuition

Word Family Frequency Level	Top 3,000 word family tokens not present in example sentences	Percentage of word family not covered
K-1 families not in input:	15	1.5%
K-2 families not in input:	100	10%
K-3 families not in input:	215	21.5%
Totals	330	11%

Discussion

The results of this study showed that experienced ESL practitioner native speaker intuition can be relied upon to create content using mostly high-frequency vocabulary since overwhelmingly the large amount of context created by native speakers fell into the high-frequency realm. In fact, in comparison to the percentage of items that fell into the high-frequency realm for the formulaic phrases alone, the addition of approximately 130,000 more tokens of example sentence context actually only reduced the percentage of tokens in the high-frequency realm by 0.92 percent (see token percentages for 1,000 word family frequency levels 1-3 in Tables 25 and 26). This copious amount of high-frequency data creation revealed that native speaker intuition can be relied upon to supply contextual content when the goal is to

create supporting context that does not add an addition learning burden in relation to the target formulaic sequence.

This study also confirms the value of a small but extremely frequent amount of word families. In total, the words used in the entire corpus of example sentences consisted of only 4,140 word families. This indicates that even when there is a great amount of data, certain high-frequency words are used repeatedly. Thus the value of high-frequency vocabulary and the collocations they occur with are confirmed. Furthermore, despite adding such a copious amount of context, only 1,517 word families were actually added since the phrases themselves consisted of 2,623 word families. Although the total tokens created resulted in a very large database, the vocabulary load (4,140 families) is feasible for learners.

One interesting aspect of this study was the style that the sentences were written in. All four native speakers wrote and used language in a subtly different style. For instance, one of the native speakers, an avid reader of fiction, more often included sentences which included quotes of what someone said in a way that is typical of fiction writing. Another more often wrote about economic issues in comparison to the other writers. Another writer, an American, created sentences involving gun violence more often that the others. It is certainly a possibility that this variety of native speakers writing sentences may have contributed to the high coverage of the top 3,000 word families of English.

Although the example sentences did cover a high percentage (89 percent, see Table 27) of the top 3,000 word families of English, why 11 percent was overlooked should be discussed as well. Ideally, writers would have included some of the words in this 11 percent in the sentences to expose learners to them. However corpora, by their nature, can never truly represent natural language perfectly. For instance, the ease with which corpora can be compiled with written texts already in digital form increases the potential for formal language to more often be included due to the nature of written texts. This is clear in how words such as *bacterium* exist within the top 3,000 words of English. Actually, the existence of the word *bacterium* in the top 3,000 word families of English is an issue, because such a word clearly has low value to learners of general English. Also, since *Vocabprofiler* utilizes word family lists partially derived from the British National Corpus, differences between British and North American English occasionally explained why these words were overlooked. A few examples found were *centimetre, flavour,*

duke, *lord,* and *pub.* Furthermore, the vast majority of the words not found in the top 34 (1,000 headword) word family lists were items that the program has trouble counting, such as word with hyphens (*middle-aged, x-ray,* etc.). Such items highlight weaknesses in the corpus or the software rather than weakness in the example sentences.

Conclusion

This experiment aimed to determine whether the intuition of experienced ESL practitioners could be relied upon to create contextual content that mostly fell into what is considered high-frequency vocabulary. Native speakers wrote nearly 160,000 tokens worth of example sentences for high-frequency formulaic sequences derived from a corpus. The resulting database was compared to the formulaic sequences alone to determine whether the content added by the native speakers mostly stayed within the high-frequency realm.

The results showed that the tokens in the sentences not only covered the vast majority of the top 3,000 word families of English (89 percent of them), 97.39 percent of the words in the sentences also fell into these top 3,000 families. Therefore, this study affirmed that native speaker intuition can be relied upon for such a task, even large-scale ones.

While this study highlighted how the intuition of experienced ESL practitioners can be relied upon to produce high-frequency contextual content, some unintended discoveries were also made. The content all four native speakers created had subtle differences in style and focus, and this variety of language may have contributed to the high coverage of high-frequency vocabulary. Therefore, future research should consider this and compare the type of language created by multiple native speakers versus only one to determine whether the subtle differences among writer styles are connected to high-frequency vocabulary coverage.

RQ9: Are there any correlations between Japanese university students' knowledge of MWUs most representative of high-frequency lemmatized concgrams and TOEFL score, item frequency or L1-L2 congruency?

Introduction

Previous research shows that it is clear that both beginner and advanced level second language learners throughout the world lack collocational fluency (Gitsaki, 1996; Nesselhauf, 2005). Researchers also point out specific aspects of certain collocations which make them have a higher learning burden than others, such as frequency and L1-L2 congruency. However, despite there being a variety of evidence that highlights these issues, to date a large-scale resource that identifies common, useful collocations deemed worthy of direct instruction by native speakers did not exist to test learners with. Although there are very large-scale lists available that are more dictionary-like (Kjellmer, 1994), these do not focus on general English as this study did, and practically speaking, it would be impossible to teach such a resource since it contains over 85,000 collocations. Furthermore, to date no large-scale studies have used the lemmatized concgramming approach utilized in this study for useful MWU identification. This current study took such an approach and created such a resource, and it was used in the following experiment to confirm and more specifically pinpoint the extent that TOEFL score, MWU frequency, and L1-L2 congruency correlate with a learner's knowledge of the MWUs identified.

Materials

This experiment utilized the same list of 11,208 high-frequency lemmatized concgrams as in research question 4, with the added MWUs identified in research question 5 and contextualized example sentences created when answering research question 8.

Procedure

First, the list of MWUs was sorted by frequency and then divided into five sections with an equal amount of MWUs in each. Then, each section was sorted by the MWU's L1-L2 congruency rating. Ten MWUs were selected from each of these five sections. An attempt was made to choose approximately five items for each of the L1-L2 congruency ratings (0-12). However, it was not possible to have an equal amount in every section because the total ratings (13 different possible scores) does not divide equally, and because of the fact that some items did not receive certain scores (none of the MWUs received a score of exactly 11, for example) and/or scores were not round numbers. However, every attempt was made to make as balanced of a sample as possible (see Table 28 below).

Table 28

L1-L2 congruency ratings of MWUs selected for testing students' collocational fluency

L1-L2 Congruency Rating	# of MWUs Selected
0	5
1.5	5
2	5
4	5
4.8	3
5	2
7	2
7.2	3
8	5
10	5
10.5	2
10.8	3
12	5

Then, a cloze test was created with these 50 items (see Appendix 25 for the test and all relevant data). Each MWU consisted of a lemmatized concgram pair (a pivot word and its collocate) of which each of the pairs were either a noun, verb, adjective or adverb, along with any other words which helped to form its most common MWU. Frequency data was collected for each pivot and collocate, and the word that was less frequent was chosen to be the target word for the questions. The rationale for this was that the less frequent item is more predicted by the more frequent pivot word. In addition, if the more frequent pivot word was chosen at the target, then there would be more of a chance that a student could guess the answer via their

knowledge of high-frequency vocabulary and not high-frequency collocation. The example sentences created by native English speakers in a previous experiment in this study were then utilized to create the cloze test items. However, for this test the aim was to create contexts that used only high-frequency vocabulary. Thus, Cobb's (2013) program *Vocabprofiler* was used to confirm that all words in the supporting context of each sentence (outside of the MWU being tested) were high-frequency (all words occurred within the first 3,000 word families of English (BNC/COCA combined)).

For example, the most common MWU representative of the lemmatized congram *line/credit* was found to be *a line of credit.* Since *credit* is less frequent than *line*, it was chosen to be the target item for production. The first letter of the target item was provided to avoid other possible answers. Thus, the following sentence was utilized:

The bank offered a line of c _ _ _ _ _ to the company to buy some new equipment.

It should be noted that in some cases (4 of the 50 questions), two letters were provided to avoid other possibilities, and in one case the first five letters of the word was provided. These modifications were determined after pilot tests were conducted first with seven native English speakers and then a group of 39 learners at the same university as those who took the final test.

The final test was then administered to as wide of a proficiency range of Japanese university students as possible. 549 students at a Japanese foreign language university campus whose student population consists of approximately 2,000 students were tested. These students were also asked to provide their TOEFL ITP scores when tested. Access to the breakdown of their TOEFL score among the language skills tested was not accessible and thus only their total score could be recorded. The results were tallied and then analyzed to determine whether their TOEFL scores correlated with their ability to produce answers on the test, and whether frequency or L1-L2 congruency played a factor in affecting their knowledge of the items.

Learners were also given the option to sign a consent form which allowed for their anonymized test data and TOEFL scores to be used for research purposes.

Finally, all of the data was analyzed to determine if there were any correlations between the variables of TOEFL score, item frequency, and L1-L2 congruency.

Results

First, all students opted to sign the consent form allowing for their anonymized test data and TOEFL scores to be used in this study. Original N-size was 549 and their results can be seen in detail in Appendix 26. Student TOEFL scores ranged from 310 to 677. The mean score was 421. A total of 14 outliers' data was removed from the study because these students did not get one question correct on the test and such data had the potential to distort the statistical analysis. Their average TOEFL score was 367. Thus, the new N-size became 535 with a mean TOEFL score of 421, low of 310, high of 677, and S.D. of 48.18. Cronbach's alpha reliability was $\alpha = .78$, and thus the test exhibited internal consistency. The highest score on the test was 52 percent correct, and the lowest was 2 percent correct. The average score on the test was 23 percent correct. Thus, it was found that the students had very low knowledge of the test items. In regards to students' TOEFL scores correlating with knowledge of the test items, the analysis did not show a correlation.

An analysis of the data was conducted to determine if there was a correlation between item knowledge and frequency level (see Table 29 below). A linear progression was not found in regards to increasing frequency versus increasing correct responses across all five levels of item frequency tested. However, if one level was removed (level 2), a linear relationship was identified which showed that as item frequency increased, so did correct responses.

Table 29

Mean scores for test items organized by frequency level

Frequency Level	M	SD	Total Correct Responses
1	0.63	0.85	338
2	1.29	1.38	689
3	0.65	0.91	350
4	1.09	0.89	583
5	2.20	1.62	1,179

Multiple regression analysis with TOEFL as the dependent variable and item frequency as the independent variable was also conducted (see Table 30 below). Due to the Bonferroni adjustment to control for Type II error, the p-value was set at .01 (.05 divided by five comparisons). This is due to the fact that multiple analyses were conducted and there is a need to lower the threshold in which we will judge the results as statistically significant because when the number of comparisons increases, so does the potential for one of them to have an outcome that is by pure chance (Davies, 2013). For the multiple regression analysis, the results were R = .57, R^2 = .33, Adjusted R^2 = .32. For the ANOVA, the results were $F(5,529) = 51.49$, p = .000.

Table 30

Multiple regression analysis and correlation coefficient with TOEFL as the dependent variable and item frequency as the independent variable

Factor	B	Beta	t	p	r
Lvl 1	8.48	.15	3.44	.001*	.42
Lvl 2	3.39	.10	2.14	.032	.41
Lvl 3	9.98	.17	3.81	.000*	.44
Lvl 4	3.54	.07	1.68	.094	.27
Lvl 5	8.05	.27	5.83	.000*	.50

The R-squared value of 32% indicates that the model explained variability of response data around its mean to an extent. It was also found that three levels of item frequency predicted TOEFL scores, and that there was a significant but small correlation between item frequency levels and item score (r = .28). The strongest predictor was level 5. The beta weight of .27 indicated that a change in level 5 item scores of one standard deviation would result in a TOEFL score increase of 13 points (.27 X 48.18). While significant predictors, item frequency levels did

not have particularly strong beta weights and thus cannot be construed as the most salient variable in predicting TOEFL scores for the sample population.

In regards to L1-L2 congruency as a factor in predicting TOEFL scores, multiple regression analysis was conducted with TOEFL score as the dependent variable and L1-L2 congruency as the independent variable (see Table 31 below). Due to the Bonferroni adjustment to control for Type II error, the p-value was set at .006 (.05 divided by nine comparisons). For the multiple regression analysis, the results were $R = .61$, $R^2 = .37$, Adjusted $R^2 = .36$. For the ANOVA, the results were $F(5,529) = 34.42$, $p = .000$.

Table 31

Multiple regression analysis and correlation coefficient with TOEFL score as the dependent variable and L1-L2 congruency as the independent variable

Factor	B	Beta	t	p	r
Cong0	11.92	.09	2.49	.013	.28
Cong1	0.39	.00	0.10	.923	.12
Cong2	14.71	.21	4.46	.000*	.50
Cong4	0.77	.02	0.51	.614	.26
Cong5	-4.93	-.05	-1.28	.200	.20
Cong7	9.20	.13	2.98	.003*	.39
Cong8	12.50	.21	5.21	.000*	.44
Cong10	6.28	.18	3.95	.000*	.46
Cong12	5.70	.08	2.00	.05	.32

The R-squared value of 36% indicates that the model explained variability of response data around its mean to an extent. TOEFL scores were predicted by four sets of congruency levels: Cong2, Cong7, Cong8, and Cong 10. Cong2 and Cong10 both had a standardized beta weight of .21, indicating that an increase in congruency scores by one standard deviation would

result in a corresponding increase in TOEFL scores by 10 points. The results generally supported the hypothesis that L1-L2 congruence generally leads to greater test scores, with a significant but very small correlation between congruency levels and overall vocabulary score (r = .06). However, the low beta weights do not indicate congruency between L1 and L2 phraseology as salient.

In summary, the results of this study show that Japanese university students could only answer 23 percent of the items on the test correctly. In regards to TOEFL scores having a correlation with collocational knowledge, there was not a correlation found. As far as frequency and L1-L2 congruency being factors that affect the learning burden of the collocations, only a small correlation was found for both variables.

Discussion

The results of this study indicated that the students have very little knowledge of the test items because the average score correct was only 23 percent. Despite the test items being a balanced selection of high-frequency collocations which native speakers have no problem producing the answers to, students still struggled with such questions. Even the highest score on the test (52 percent correct) would be considered as failing by standard measures in Japan as 60 percent is the standard passing grade. This is not surprising since comprehensive resources to teach such items does not yet exist and therefore students are not being taught such knowledge directly. Because such resources do not exist, textbook writers have no resource to reference to when selecting items to focus on. Thus, collocations are not directly taught and the obvious end result is a lack of collocational fluency.

The results of this study also indicated that there was not a correlation between TOEFL scores and the test items. There are a number of reasons why this may be the case. First of all, if students had taken the TOEFL iBT test which requires speaking and writing and not the TOEFL ITP test, then the data may have correlated because the test utilized in this study required productive knowledge. In addition, the lack of any comprehensive resources that identify high-frequency collocations could also play a role. Since no such resource currently exists, it is not possible for proficiency test creators to load on such items.

Previous studies have shown that frequency and L1-L2 congruency play a factor in increasing an item's learning burden. This study, however, was not able to show a strong correlation between frequency and L1-L2 congruency and the students' ability to provide a correct answer. This is because the students' mean scores were so low that a proper analysis was not possible. A lack of collocational fluency across the board makes it impossible to extract the data necessary to show a correlation. Thus, frequency and L1-L2 congruency may still be factors and that further research should be conducted to help make the extent to which they are more salient. For example, a test of receptive knowledge may provide data which indicates a stronger correlation. Regardless, the overall lack of collocational fluency found in this study indicates that this aspect of vocabulary depth knowledge needs to be focused on more by students, teachers and materials writers.

Conclusion

This experiment examined Japanese university students' knowledge of high-frequency collocations. It found that their knowledge of the items tested was extremely low with an average of only 23 percent correct compared to native speakers, who in test piloting got perfect scores. This study also found that TOEFL scores did not correlate with collocational knowledge, and that there was only a small correlation between the factors of frequency and L1-L2 congruency and collocational knowledge. The students overall lack of knowledge, even students with high TOEFL scores, limited this study's ability to show a correlation between frequency and L1-L2 congruency and collocational knowledge, and thus more research is needed to determine just how much of an influence these two factors truly have on affecting a collocation's learning burden. However, this overall lack of knowledge does clearly indicate that this is an area that needs much more focus by teachers and materials writers to help Japanese university students achieve full fluency in English.

Research Questions and Answers Summary

1. What is a frequency data cut-off for lemmatized concgrams that results in a list consisting of 2-3,000 word families?

One occurrence per million tokens proved to be an ideal frequency cut-off for this study. It resulted in a list of items that could be practically taught (11,208 lemma pairs). This number seems impractical for explicit instruction at first glance, but in reality these MWUs only consist of approximately 3,000 word families in total, are mostly high-frequency, and have high coverage of the top 3,000 word families of English. Thus, this frequency cut-off was determined to be ideal and practical.

2. To what extent is corpus dispersion data useful for identifying MWUs that are deemed worthy of instruction by native English speaker intuition?

This experiment revealed that the type of data and methodology used was not useful in identifying MWUs most representative of high-frequency lemmatized concgrams. Various parameters were experimented with, and their results were judged by native speaker intuition to be too inclusive or too exclusive. Items that were considered by native speakers to occur across a wide variety of texts and thus had value in explicit learning was excluded by some parameters, while other parameters marked items as being balanced while native speakers viewed such items as not having value in explicit instruction. Thus, it was determined that a combination of manual checking using native speaker intuition and a corpus data analysis such as the one used in this experiment, while time-consuming and subjective, was preferable in comparison to the steps taken in this experiment.

3. To what extent is corpus chronological data useful for identifying MWUs that are deemed worthy of instruction by native English speaker intuition?

This experiment revealed that the type of data and methodology used was not useful in identifying MWUs most representative of high-frequency lemmatized concgrams. Various parameters were experimented with, and their results were judged by native speaker intuition to be too inclusive or too exclusive. Items that were considered by native speakers to be dated, only occurring during a limited time period, or too modern and thus not yet established were excluded by some parameters, while other parameters marked items as having balanced chronological data distribution while native speakers viewed such items as not having value in explicit instruction. Thus, it was determined that a combination of manual checking using native

speaker intuition and a corpus data analysis such as the one used in this experiment, while time-consuming and subjective, was preferable in comparison to the steps taken in this experiment.

4. To what extent is consideration for colligation an important criterion for identifying MWUs that are deemed worthy of instruction by native English speaker intuition?

This experiment revealed a number of things. First, consideration for colligation can improve results of MWU identification. However, the amount of items that were actually improved upon in the current study was a very small percentage of the total. In addition, no dedicated software existed, and thus a very complex, cumbersome, and time consuming methodology was required. Thus, while colligation can sometimes be an issue it is not necessarily a significant issue for the items in question. This experiment also revealed that methodological and software improvements are certainly needed as well in regards to analyzing data for colligation.

5. What percentage of MWUs is deemed by experienced native speakers who teach English as a second language at university worthy of expanding beyond their most frequent exemplar to provide learners with useful information about how the items commonly occur formulaically?

This experiment highlighted the value of extending MWUs beyond the core pivot and collocate. Over half of the items examined during this experiment were deemed by native speakers to be worthy of expanding beyond their most frequent exemplar. While very time consuming since it must be done by native speakers manually and finding a technological solution is not an option, this type of data analysis was still deemed to be an essential step in a study such as this.

6. To what extent is semantic transparency an important criterion to consider when attempting to identify collocations deemed worthy of direct instruction by native speakers?

Only 14 percent of the items examined were considered to be either semi-figurative, figurative, a core idiom, or had features that prevented them from easily being understood (such

as when a much less common homonym is part of the MWU). These results contrast sharply with how some researchers insist that literal collocations not be taught explicitly. The vast majority (86 percent) of how high-frequency vocabulary collocate in the form of MWUs are actually literal, and thus if they are excluded from direct instruction, how can such vocabulary be taught? Thus, these results highlight a new perspective on how high-frequency vocabulary collocate and what should be considered worthy of explicit instruction.

7. To what extent is L1-L2 congruency an important criterion to consider when attempting to identify English MWUs deemed worthy of direct instruction by native speakers to Japanese learners?

56.4 percent of the items examined in this experiment were deemed to be incongruent to an extent with the L1 in question. More specifically, on a scale of 0-12 with 0 being fully incongruent, 997 items received a rating of 0-3, 2,419 items received a rating of 4-6, 2,905 items received a rating of 7-9, and 4,888 items received a rating of 10-12. This high percentage of items being incongruent highlights the importance of conducting L1-L2 congruency analysis because that means that half of the items will have a high learning burden that the other half. Such items deserve extra teaching time, and with it, learners can avoid typical errors that are derived from their L1 influence.

8. To what extent is native speaker intuition useful in regards to high-frequency vocabulary usage in context creation?

This experiment revealed that native speaker intuition is very useful when the task is to create context using high-frequency vocabulary for the MWUs in question. Native speakers wrote nearly 160,000 tokens of content to create an example sentence for each of 11,208 MWUs. An analysis of the added content revealed that not only covered the vast majority of the top 3,000 word families of English (90 percent of them), 97.39 percent of the words in the sentences also fell into these top 3,000 families. Therefore, this study affirmed that native speaker intuition can be relied upon for such a task, even a large-scale one.

9. Are there any correlations between Japanese university students' knowledge of MWUs most representative of high-frequency lemmatized concgrams and TOEFL score, item frequency or L1-L2 congruency?

This experiment examined Japanese university students' knowledge of high-frequency collocations. It found that their knowledge of such items was extremely low. This study also found that TOEFL scores did not correlate with collocational knowledge, and that there was only a small correlation between the factors of frequency and L1-L2 congruency and collocational knowledge. Since previous research indicates that L1-L2 congruency should have an effect on collocational knowledge, the results of this study contradict such findings. However, it is possible that because knowledge of the test items was so low, the effect of L1-L2 congruency was not even registering in the data. Similarly, one would expect a student's collocational knowledge to increase as his/her proficiency (as measured by TOEFL) increases. However, this study did not show such a correlation either. Again, it is possible that because knowledge of the test items were so low, such a correlation could not be found in the data. It also may be possible that TOEFL results do not reflect collocational knowledge. Thus, more research needs to be done in regards to these two points.

Conclusion

This chapter introduced and discussed the scope of the research questions, and then gave detailed descriptions of the materials, procedures, and results of the attempts to answer them. The findings of each answer were also discussed in detail as well, and a summary of each research question and their answers was also provided.

Chapter 5
Implications and Applications

Introduction

The journey to answer the questions set forth in this thesis led to a number of significant discoveries, methodologies being developed, resource creation, and rethinking of theories. These were not only the answers to the questions set forth, but also the revealing of issues that were not yet salient when this research began. The path to answer the research questions led to

the development of methodologies which had not existed, and can now be used by future researchers to make further discoveries in the field. Furthermore, some answers that were found revealed that certain methodologies did not produce useful results in comparison to the time invested in them. This information can also be used by future researchers to avoid certain paths that do not produce fruitful results.

Answering the research questions in this thesis also led to the creation of a major resource which to date had not existed, but has been called for by researchers for a number of years. This resource has the potential be used to improve upon the efficacy of second language acquisition through utilizing it for direct instruction or materials development. Many questions still remain, such as what is the best way to study such materials, but despite this, the first step towards identifying such items has at least been taken. This resource and the methodologies that were used to create it can of course be improved upon. However, at least practitioners and researchers in the field have something to now work with.

This current research also resulted in new questions being posed. For example, the discoveries made can lead to a reconsideration of what exactly should and should not be considered a collocation and what difficulties identifying such items actually entails. These contributions to the general theory of word co-occurrence should prove valuable for researchers when thinking in more general terms about what language is and how it should be best taught.

Unexpected discoveries

This study resulted in a number of unexpected discoveries as it progressed down the path to answering the research questions it set forth.

First, as the literature review was being conducted it was noted that numerous researchers pointed out how answering such questions as in this thesis would be highly beneficial to learners, but such questions had yet to be answered. However, when the experiments in this study were undertaken it became clear why this was the case. The research questions could not be answered without help from numerous volunteers to accomplish many of the tasks. In addition to simple but time-consuming data collection and analysis, the need for highly trained translators was also required. Putting together such a team of volunteers and keeping them motivated to continue the work for a number of years was challenging. Thus, it became clear that this was work that one

researcher could not do alone. However, because such volunteers were able to contribute their time and expertise, this thesis' goals were accomplished to the best extent possible under practical limitations. This is useful information for any future researchers as well because there are still many unanswered questions, and to answer them will probably require a similar collaborative effort.

This research project began with some assumptions that did not prove true as well. It was originally believed a methodology which analyzed the dispersion and chronological data available from the COCA to help identify only items with balanced distribution could be created. However, despite utilizing a variety of parameters, all proved to be not useful in accurately achieving the goal set forth. Therefore, manual checking using native speaker intuition became necessary. This was not expected, and added a significant amount of work to the study. Furthermore, the need for an analysis of the dispersion breakdown of the COCA and how that possibly could be improved upon became clear. There was also some other discoveries regarding the make-up of the COCA itself. For example, a large amount of recipe-related language in certain sections of the corpus was not expected, and thus steps needed to be taken to deal with such issues.

In addition, the realization that the vast majority of the MWUs were judged to be semantically transparent was also surprising. There was an expectation that there would be more semi-figurative and figurative formulations, but in reality, the way that high-frequency vocabulary collocate is mostly in literal formulations.

A satisfying discovery was the reliability of native speaker intuition. Certain tasks in this study required the use of native speaker intuition for judgment and also for content creation. The ability of native speakers to utilize high-frequency vocabulary when creating example sentences for each of the MWUs identified ended up being more useful than expected since the very large amount of added content via the example sentences they created only ended up having approximately one percent of the vocabulary being outside of what is considered to be high-frequency. This discovery has a variety of implications for future research, and not just for research regarding collocations.

Another positive discovery was the word family breakdown of the final resource. When the findings of this study was discussed with other researchers and teachers, a similar reaction

occurred when it was mentioned that this thesis identified approximately 11,000 MWUs that should be taught explicitly. For these researchers and teachers, the volume is impractical for direct instruction. However, when it was pointed out that these 11,000 MWUs only consist of only approximately 3,000 word families, the realization that learners such items was possible occurred. In fact, it is not the equivalent of studying 11,000 new items. Rather, it is studying the way that 3,000 different items combine with each other in various ways. For instance, *run faster, take a walk,* and *moved away from* all occur in the list. If a learner masters these items at one point in a list, and later is exposed to *walk away* and *run away* in the same list, they are not learning any new vocabulary whatsoever. Viewing the resource from this perspective makes it possible to realize that the learning burden is actually not impractical.

Development of methodologies

Researchers agree about the value of high-frequency vocabulary and learning how they collocate to achieve second language fluency, but a large, comprehensive (but not dictionary-like) resource has yet to have been created. The goal of this current research was to create such a resource. Early on in the research, it was clear why this resource did not exist yet. Not only was the work so extensive and time consuming, methodologies which identified the way language naturally occurs did not yet exist for each of the steps that needed to be taken. Furthermore, some methodologies already existed, but needed to be improved upon.

Early on in this study, it became evident that new methodologies were going to be needed. Researchers have been talking about collocations and their importance for many years, but only in recent years has technology and theory been improved to the point where researchers have begun to talk less about *collocations* and more about *concgrams,* and particularly for this study's goal of identifying the high-frequency collocations of high-frequency vocabulary, *lemmatized concgrams.* In fact, no other study that aimed to identify high-frequency collocations has taken such an approach so it was not a surprise to discover that a number of new methodologies needed to be invented.

First, how much data is necessary when we try to identify the MWU most representative of a lemmatized concgram? This study showed that 500 instances of co-occurrence produced similar results as 1,000 did. Next, there was the development of a methodology to grapple with

colligational issues. However, a methodology which gave guidance as to which words should be searched and replaced with colligational markers did not exist. Examining the data showed that certain categories seemed to benefit from such a procedure, such as pronouns, months, days of the week, ordinal numbers, and cardinal numbers. Unfortunately, dedicated software to replace such words with colligational markers does not exist and thus this study utilized multiple pieces of existing software that had such capability. In the end, after a very complex methodology was created, this task was achieved.

After that was accomplished, additional steps needed to be taken to delimit items identified to only those which have high value for students to learn. For example, in the past researchers have utilized certain criteria, such as dispersion, to help identify only the most useful items for students to learn. However, to date, no research had examined the chronological data of collocations. Certainly, it will not be beneficial for students to learn either very dated collocations, collocations only occurring within a short time frame (trend-like or time-sensitive event related occurrences), or collocations which have yet to be established (new items which occur in high-frequency but have yet to be confirmed as permanent parts of the language). Thus, this study created a methodology to examine such corpus data to determine whether or not this criterion could improve the resource as a whole. While this methodology did identify some items that were not of value to be learned, these items were in very small numbers. Furthermore, it was determined that, even though a variety of parameters were experimented with, computer data analysis itself was found to not be useful. Often, the parameters would either be too inclusive (flagging items as having unbalanced chronological distribution which were actually of value to learn) or not inclusive enough (not flagging items that did have unbalanced chronological distribution). Thus, it was determined that a manual examination using native-speaker intuition was essential, and that even with this, the number of items identified was so small that if the study was a large-scale one (as was this study) then it may not be worth the effort.

In addition, software that could identify the MWU most representative of a lemmatized concgram also did not exist. Concordance software does exist, and such a task can be done with it, but not without *noise*. For example, if a mini-corpus that contains 500 instances of co-occurrence of the lemma *take* and *walk* is examined with currently available concordance

software, the first most common MWU identified will not be *take a walk* (which is actually the most common MWU these two lemma both occur in), but rather *of the, and the*, and so on. Only after much so-called *noise* will we eventually begin to see instances of where *take* and *walk* co-occur in MWUs. Removing such noise is extremely time-consuming, and thus software was needed that could focus only on searching for MWUs that only contained both lemma. This was a complex task because we are not just searching for *take* and *walk*, but rather we are search for *take, took, taking, takes, walk,* and *walks*, and only when both separate lemma co-occur. It was quite clear that professional help was needed. The author of the most downloaded concordance software in the world (*AncConc*), Waseda University Professor Laurence Anthony, was contacted and the discussion of how such a complex task could be achieved began. After a year of planning and software development and testing, the creation of such software was finally achieved. First, a complete lemma list was needed. This existed, but it needed to be modified to deal with homonyms because such instances would prevent the software from functioning properly. Then, the software also needed to be able to process files in bulk because this study was examining over 11,000 lemmatized concgrams and processing 11,000 files manually would be too time-consuming. A number of technical barriers were discovered and traversed, and in the end after many trials and tribulations, the MWUs were finally identified. Such software is now available upon request from Professor Anthony.

However, technological solutions were not the only way this study contributed to new methodologies. It was discovered that in reality, technology could only take one so far, and that, at some point, if quality results that could practically be used for teaching was the goal, native speakers manually analyzing the data was essential. This was discovered when the MWUs identified by the software were examined. Often, the most frequent MWU occurring was not the best choice for teaching and an extension of it was more ideal. For example, the MWU *come to terms* was identified as the top MWU occurring for the lemma pair *come* and *term* at 243 out of 500 instances. However, second in rank was *come to terms with* at 229 out of 500 instances. When native speaker intuition is relied upon, it is clear that in such instances, it is preferable to extend the core top MWU by adding *with.* In fact, this methodology proved to be extremely fruitful in that native speakers opted to do this in more than half of the items examined.

However, when we rely upon native speaker intuition we are introducing a subjective element into the data analysis. Can this truly be relied upon as a methodology? This study revealed that the answer to that question was to a practical extent, yes. The aim of this study was to create a practical resource that could be used to teach high-frequency collocations. However, just teaching these collocations in the form of the MWUs most representative of them is not enough. Providing a full example sentence helps students learn limitations and/or appropriateness of usage of these items. Therefore, an example sentence was created for each of the approximately 11,000 MWUs identified. Care needs to be taken when creating such supporting context in that the context should not result in an increase in learning burden. In other words, the context added should not contain vocabulary that is of a higher learning burden than those within the MWU itself. Ideally, all supporting context should be high-frequency vocabulary. So, this led to the question of whether or not native speaker intuition could be relied upon to create such context. Despite adding of over 130,000 words of context by native speakers, the amount of words that could be considered as "high-frequency" (words that were within the top 3,000 word families) was actually very high at 98.2 percent. Therefore, this study confirmed that such a methodology was useful.

Furthermore, it should be noted that despite the use of a large corpus (COCA) compared to previous research done using smaller corpora (BNC), there were still instances of weaknesses in the data that computer analysis could not grapple with, thus again highlighting the need for manual checking of data. For high-frequency collocation selection, Ackermann and Chen (2013) also found a manual checking and vetting of items necessary in addition to what results their corpus data analysis could produce. A corpus itself contains natural language, but it is not a mirror reflection of language as a whole. The compiler of the corpus may attempt to include data from as wide of resources as possible in the most balanced way possible, but it will never unequivocally emulate natural language. Therefore, the existence of noise in the data is inevitable, and this study highlighted the types of noise that can appear and the extent to which it can affect the quality of any resource that is derived from such corpus data. Overall, the results point to the COCA as being a very useful resource that has very minor flaws. For example, language related to recipes had higher than would be expected frequency counts in the corpus because certain magazines that it sourced data from contains recipes. Recipes have a tendency to

have certain language, such as the collocates *cup/sugar,* repeat more often than they truly occur in natural language, and thus such language ended up being identified as "high-frequency" before dispersion was considered. So, if one simply relies solely upon a corpus' computer data, then, depending on the research's goal, results do have the potential to be atypical in comparison to natural language. So, this points to the importance of a methodology including the manual checking of data for such weaknesses.

Another example of how this study contributed to a methodology can be seen in the experiment concerning semantic transparency. Researchers agree that semantic transparency affects a collocation's learning burden. Thus, when collocations are examined, they are broken down into categories such as *literals, semi-figuratives, figuratives,* and *core idioms*. One step that was taken in this study was to examine the nature of high-frequency lemmatized concgrams and determine what percentage of them was in each of these semantic categories. However, when the items were examined to place them into one of these four categories, some items did not seem to fit. A new *outlier* category was thus created for items which had the potential to have a higher learning burden than a literal formulation, but did not meet the criteria to be placed in the other categories. For example, when a MWU contained a homonym that could be easily misunderstood (when the significantly rarer homonym is used), such as *bear* in *bear children,* such items were put into this outlier category. Collocations were also given this rating when they had very specific meanings which learners have a high probability of misunderstanding (e.g., *boot camp, social security*, *foster care*). In addition, if a collocation seemed to be formed arbitrarily (there is no rhyme or reason why a particular word is used, and not another logical alternative), it was also given this rating. Examples include *take measures, deliver a speech,* and *to stand trial.* For instance, why do we *take* measures and not say *create measures*? Why do we *deliver a speech* but don't *deliver gossip*? Furthermore, would it not be more logical to just say *have a trial*? Recognizing these arbitrary ways in which language combines is essential to recognizing learning burden. So, by examining the items for this criterion of semantic transparency the potential for a new category was discovered. Certainly more research needs to be done in regards to this in the future, but regardless, this discovery has the potential to improve upon future methodologies regarding semantic transparency.

Yet another methodology that resulted from this study was in regards to L1-L2 congruency. Researchers agree that L1-L2 congruency affects a word or phrase's learning burden. One-to-one congruency equates to a lower learning burden, while when word or phrase is said in a totally different way between two languages, such items will be much more difficult to learn. Additional time needs to be spent on such items because of this higher learning burden, and by identifying them, teachers can focus on them to help students avoid errors, such as in production (direct translation from the student's L1). However, despite researchers being aware of this issue, no methodology to specifically compare and rate L1-L2 congruency between Japanese and English and deal with all the particular differences between these two languages existed when this current study began. This study specifically examined congruency between the English MWUs identified and their Japanese translations. A point scale was created which gave each word in a MWU a certain score. A number of issues arose that also needed to be dealt with. For instance, when linguistic phenomenon did not exist in the L1 (for example, English articles (*a/the*) do not exist in Japanese), such words were not including in the rating. Furthermore, a point system had to be devised when a word and its translation were in the same word family but a different part of speech. Similarly, a rule was created for when words had only slight semantic differences. A number of other rules also were created to deal with a variety of issues that arise when such a complex comparison of two languages is conducted. In all, these steps highlight the complexity of conducting L1-L2 congruency comparisons since other languages will obviously have other differences that need to be dealt with in special ways. Therefore, this first step towards a methodology for conducting L1-L2 congruency analysis is certainly a valuable contribution to the field.

Creation of a resource

As was mentioned in the previous section, a large-scale (but still having potential to be explicitly taught) resource which identified high-frequency collocations of general English did not exist, and therefore the creation of a number of new methodologies was necessary. These methodologies led to fruitful results in that the resulting list has the potential to be of high value to learners and practitioners of ESL. In general, the core English version of the list has value for learners across the globe. Currently, there are Japanese translations of all MWUs and example

sentences, but translations and further L1-L2 analysis could be conducted for any language in the future. Therefore, this resource could be considered as just beginning its development for students across the globe rather than being considered as complete. Further testing with more students and at varying years of university, at various universities throughout Japan and across the globe are all certainly called for.

This resource also has high potential to be used as a reference when creating other materials, such as textbooks or educational software. The reality is that most textbooks still focus on teaching isolated vocabulary, which is quite an inefficient and unnatural way to learn a language. Yes, collocations and MWUs do exist in textbooks, but often these books are not bringing students' attention to them (Gitsaki, 1996). Previous research has shown that when a learner's attention is not brought to them, the learner is not able to notice them. Thus, textbooks really need to point them out. However, materials writers have been lacking a resource to reference when choosing such items. For many years, materials writers have relied upon comprehensive high-frequency vocabulary lists, but until now, they have not had access to a resource that identified the MWUs most representative of how high-frequency vocabulary collocate.

In addition, this resource also has the potential for direct study/explicit instruction. Just as students have studied word lists in the past, students can now study such words, but also with their collocates within MWUs. Although rote learning and L1 contrastive analysis has been dismissed by some as a relic of the past, more and more researchers are now reconsidering its value because of its high efficiency and ability to be structured in an organized way (Avery & Baker, 1997; Hopkins & Bean, 1999; Rodriguez & Sadoski, 2000). So, with the existence of this resource along with L1 translations, such study is now possible for a particular learner group (Japanese learners). Currently, this resource is available for download on Apple and Android smartphones and tablets in the form of a Leitner-algorithm based flashcard app called 英語マスター1万 [English Master 10,000]. This app can also be found by alternatively searching for this thesis' author's name on the online app stores.

There is also the potential for other resources to be created as well. For instance, Cobb's (2013) *VocabProfile* has the ability to highlight the frequency of isolated words in any text by simply inputting such text into its online interface. However, now that the resource this research

created exists, a similar type of software could be created for high-frequency MWUs as well. This would be very useful for determining which high-frequency MWUs are present in texts that already exist. By bringing students' attention to them, this will enable them to notice and master such knowledge, since previous research indicates that this does not happen implicitly. This current resource can serve as a basis for created such additional resources.

Contribution to theory

This current research has led to a number of realizations in regards to the theoretical knowledge understanding collocations that were not salient before. First, the data point to the lexical approach towards identifying collocations as the most useful for the goals of this study. This study combined that approach along with some aspects of the structural approach with its steps taken to deal with colligation. These steps towards identifying items whose results could be improved upon through a colligational treatment only proved useful to a small extent. While certain items were improved upon, the numbers paled in comparison to those which did not benefit from such a treatment. Therefore, the data in this study points to a lexical approach being the most advantageous. When the lexical approach is accepted, it is then possible to begin to think of language in a very different way. To quote Lewis (1993), language "consists of grammaticalised lexis, not lexicalized grammar" (p. vi). In other words, lexis organizes language, not grammar. However, the improvements that were made through the colligational treatment taken in this study makes it clear that black and white condemnations or acceptance of one theory or the other is not appropriate. Moreover, it is also possible to point out features in English that even support the semantic approach to understanding collocations. In fact, all three approaches, structural, semantic, and lexical, are valid and it simply depends on what the goal of the study is when choosing which approach to take. For the current study's goal, the lexical approach supported somewhat by the structural approach proved to be the most appropriate.

In more general terms, this study contributed to the theory of collocations by helping define what is or is not a collocation. Just as with the approach one takes toward understanding collocations, there are a variety of equally valid definitions of what is a collocation, with some being more inclusive than others. However, from the perspective of the ESL practitioner and/or learner, and with the very practical goal of identifying which frequently co-occurring language

features should be taught explicitly to help learners attain second language fluency, this study came to some particular revealing conclusions.

Specifically, when the way high-frequency vocabulary collocate is examined and the MWUs most representative of such collocations are identified, it becomes evident that the vast majority of such items are literal formulations. However, a number of researchers do not consider such formulations to be 'collocations' and also do not believe they deserve explicit instruction because semi-figurative, figurative, and idiomatic formulations have a much higher learning burden. There is nothing wrong with such a view in that it is a logical and appropriate view of a particular type of linguistic phenomenon. However, the goal of this current study was not to describe or define linguistic phenomenon in rigid ways. Rather, the goal was to identify the way high-frequency vocabulary co-occur to help learners master how to use it properly, and in turn, attain fluency in that area of second language proficiency. So, the real question is not what is or is not a collocation, but rather what commonly co-occurring language needs to be taught?

For example, if the approach of not accepting any literal formulations to be collocations is taken, as researchers such as Moon (1994; 1997) believe, it is very problematic in that there is a major loss as to the volume of how high-frequency vocabulary collocate since only a small minority of the items identified in this study are non-literal formulations. The idea that a learner could gain 'collocational fluency' by simply mastering the 1,000 or so non-literal formulations identified in this study is unfeasible since mastering such knowledge often takes learners a lifetime. In fact, a number of researchers have already pointed out that native speakers can have upwards of hundreds of thousands of collocations in their lexicons. In addition to that issue, there is another to consider: L1 congruency. Imagine that a literal formulation is excluded as being considered as a collocation, but that formulation is incongruent with the learner's L1. In that case, the learner will have a high probability of making an error with such an item. For example, when they try to produce the formulation, they may directly translate how it is said, word for word, from their L1 and thus create an unnatural formulation. In other words, they will make an error. Isn't this exactly what the task of the teacher is, to help students avoid errors and to help them produce accurate language? If that is true, then as a teacher aiming to create a resource for real students that have real goals and needs, there is an obvious need to focus more

on meeting those needs rather than limiting myself to describing linguistic phenomenon in a rigid way. This current research has helped me to discover that this flexibility in theory is absolutely essential if one is to conduct such 'applied' linguistic research as this study aimed to accomplish.

Considering the concepts mentioned above, it becomes clear that understanding how to solve the problems at hand is more about perspective. When looked at from this perspective of 'what needs to be taught' rather than 'what falls into this rigid categorization of collocations', one begins to think very differently about collocations and how to develop such fluency in second language learners. In fact, as a teacher and researcher, I have come to care less about what is a collocation or not. I am rather more interested in what is or is not a lemmatized concgram, its frequency in a corpus, whether or not that it has balanced dispersion, whether or not that concgram would benefit from a colligational treatment, whether or not that concgram should be extended beyond its core unit, whether or not it is semantically transparent or not, whether it is congruent with the learner's L1 or not, and what the target learners' knowledge of it is.

Conclusion

This chapter discussed the implications and applications of this study. It described unexpected discoveries, such as the need for assistance because data analyzation was so time consuming. It also discussed the development of new methodologies, such as the creation of new software to accomplish the task of extracting MWUs from mini corpora of lemmatized concgrams. This chapter also mentioned how this study has led to the creation of a resource, a list of MWUs that have been translated into Japanese, and the potential for it to be translated into other languages in the future. In addition, contribution to theory was also discussed, such as how only considering non-literal formulations as 'collocations' is problematic for the goal of this study in that it would exclude the vast majority of high-frequency formulations.

Chapter 6
Concluding Remarks

Introduction

This chapter will explain a variety of limitations that this study possesses. It will also discuss potential directions that future research should be done in. Not only has this current study opened up new paths for future research, difficulties found when attempting to answer some of the research questions highlighted how some questions could not be answered to the extent desired and that more research will be necessary to fully answer these questions. Finally, this chapter will conclude the study with an overview of this study's objectives and rationale, approach, results, and final thoughts.

Limitations

As mentioned previously, although this current research possesses a number of clear limitations as to how its results can be interpreted, used and relied upon, it should still serve as a good first step towards achieving the goal of helping learners master collocational fluency. But why do these limitations exist? As stated in the research paradigm section of this thesis, the overarching approach taken to achieve the goals of this research was post-positivist. In other words, steps were taken to answer the questions that approximated reality while acknowledging unavoidable weaknesses.

For example, practically speaking, high-frequency vocabulary lists such as Nation's (2004) *BNC 3,000* or West's (1953) *GSL* are very useful resources that have been used to achieve practical learning goals for years. Certainly some of the words in the lists can be improved upon. Certainly one could also argue from certain perspectives that some of the excluded vocabulary that ranked between entries 3,001 to 3,100 are more useful than items that ranked from 2,900 to 3,000 in Nation's list. However, that is beyond the point. There is a variety of perspectives that one can take to approach such a goal as in this study, and all are valid, but when one answer is needed to make the creation of some sort of resource to fill a gap possible, some decisions must be made that make the results unavoidably limited in their, for lack of a better term, 'validity'. As I progressed along the journey of solving the research questions set forth in this thesis, I found more and more that the key to solving the task at hand was to avoid harsh black and white thinking, admit that results will never be unequivocal, and make the best approximation possible within unavoidable constraints.

With that said, there are a number of specific limitations that should be acknowledged when interpreting the results of this study. First, just as Nation (2001) said that setting a frequency cut-off is unavoidably "arbitrary" when speaking of high-frequency vocabulary list creation, the same is true for this study. Are there useful collocations that occur less than the one occurrence per million tokens that this study took? Yes there are. So, with this in mind, this is not a comprehensive unequivocal list. Furthermore, could this frequency cut-off be too inclusive and thus identify items whose value to study explicitly be questioned? The answer to that question is yes as well. This study created a large-scale resource that contains over 11,000 MWUs. Certainly if a small-scale study was done, the data could be more precise but because of this scale, this frequency cut-off reliability limitation needs to be acknowledged.

In addition, although the bulk of this study was quantitative, the realization that corpus data could not be solely relied upon to produce results which agree with native-speaker intuition, there was a necessity to include subjective judgments by native speakers to help achieve the goals set forth in this thesis. This included judgments as to what MWUs should and/or should be considered as being useful across a wide variety of topics (balanced dispersion), which were chronologically stable, which MWUs would students benefit from when having their MWU core extended (e.g., *come to terms* versus *come to terms with*), and regarding the MWUs' semantic transparency. Furthermore, unavoidably subjective judgments were also made when conducting L1-L2 congruency analysis. This is in addition to the translation of nearly 160,000 words of content into an L1 (Japanese). Translation is not an exact science since there are a variety of ways that something can be translated. However, the translation team did its best to examine the example sentence, determine the exact meaning conveyed by the MWU, and to translate it into the best equivalent natural Japanese possible. L1-L2 congruency analysis is not an exact science either. In fact, this study itself had to create an original methodology just to conduct it. It should be noted, however, that due to the extremely time-consuming process necessary and the difficulty in finding volunteers qualified enough to analyze the data for this criterion, how this study's results can be interpreted has clear limitations since the difficulty of the task could have been compensated for by increasing the number of respondents. Unfortunately this was not practically possible and thus more work should continue in regard to this in the future.

Moreover, some aspects of this study do also have the clear limitation of only being applicable to Japanese learners. L1 congruency analysis was only conducted with Japanese. Obviously the results will be entirely different depending on the L1 in question. Thus, that aspect of this study cannot be applied to any other group of learners. Conducting L1 congruency analysis with other L1s is clearly preferable, but because of the scope of this study that was simply not practically possible. In addition, this study's finding in regards to students' knowledge of the items was also limited to Japanese university students. Thus, learners from different levels of education or different backgrounds may have varying results if tested in the same way.

Furthermore, this study also has limitations in regards to interpreting its results regarding Japanese university students' knowledge of the MWUs identified. This study shows that their knowledge of the items identified was extremely low, so low in fact that the data did not even correlate with TOEFL scores. However, this study only tested productive knowledge. But what of receptive knowledge? Could the results be very different from that of productive knowledge? They certainly can. Thus, more research needs to be done to determine the extent to which Japanese students have knowledge of such items. Furthermore, this is making the assumption that these items are important and worthy of study. That has yet to be determined. In fact, that is the most important limitation to acknowledge with this study. In general, this study does not try to refute or profess any particular belief. It was simply an exercise attempting to identify certain items that are worthy of explicit instruction for the purpose of improving upon collocational fluency. It laid out a methodology and showed the results of that methodology. If practitioners find value in these results and use them, then that is an added bonus. However, again, this study merely laid out a methodology that could be used to achieve the goals it set forth. One should note that the words "could be used" is used, and not "should be used". Data to support why certain decisions and methodologies were taken was provided, but this study does not claim that other methods could not produce better results. This study simply shows that these are the results when these particular steps are taken. Thus, there is a tremendous amount of future research that is still needed.

In regards to the reliability of the data set utilized in this study from the COCA, some limitations were discovered that should be pointed out. A few years after all the experiments

were completed, the frequency data of the lemma was double-checked and were found to be slightly off. However, this difference of frequency data is consistent. The total frequencies of lemma pairs in the data set Word List Plus Collocates was thought to be from a completed section of the COCA (1990-2009). This would lend to easy replicability and consistency when researchers use the list in conjunction with the COCA's interface. However, the data is consistently slightly less. For instance, the lemma pair *out/cigarette* has a frequency count of 862 in Word List Plus Collocates, while it has a larger count of 873 when data from 1990-2009 is tallied. So, this minor inconsistency seems to mean that Word List Plus Collocates was compiled slightly before the 2009 data set was completed. However, when the creator and administrator of the COCA (Professor Mark Davies) was contacted about this inconsistency, he explained that Word List Plus Collocates was compiled using data up until April 2010. If this was the case, then frequency totals in Word List Plus Collocates should be slightly higher than that of the data set from 1990-2009. However, they consistently are not. Other lemma pairs were searched for, and a very similar small percentage of difference (0.2 percent) was found. Professor Davies could not explain why this is, but the difference is so minimal that it is not considered to be an issue for this study's results or replicability.

In addition, total frequency counts of the final list in this study are also consistently off by about 5 percent in comparison to frequencies from Word List Plus Collocates and 1990-2009 data. The totals in this study's final list are always higher by around five percent. These frequency counts were extracted by the COCA's online interface from the 1990-2009 section over four years ago. There is no explanation for this difference, but since it is consistent it seems to mean that at some point how the COCA counts data or its data set may have been modified slightly.

Furthermore, there are very slight inconsistencies in regards to duplicate entries and which data was kept. For instance, when a search with the COCA's interface for the lemma pair *figure/out* is conducted, the total frequency count is 28,076. However, when the reverse search is conducted (*out/figure*), the total frequency count is 28,075. This is merely a technological limitation of the COCA's concordance program and how it searches for data. Professor Davies was contacted about this inconsistency, and he explained that some entries in the data set are not counted if they cannot be separated by a period when the program analyzes the data

grammatically (*out.* or *figure.*, for example), but this is only done for the pivot word and not the collocate that is searched for. Thus, when reverse searches are conducted the numbers may be slightly off. Again, these differences are extremely minor and should not pose as a significant issue for interpreting this study's results or for its replication.

Since the COCA is such a large resource and its maintenance must be extremely complex and time consuming, it is not surprising that this is the case and that Professor Davies could not explain the slight differences. However, they should be pointed out. In regards to replicability of this study or interpreting its results, these differences only amount to a small issue because of the large-scale of the study. Whether an item on the list has a frequency count of 873 or 862 does not really pose an issue because items on the list are not separated into sections where such differences are an issue. It is true that some items may not be included in future studies because of frequency differences at the cut-off of 500 occurrences, however it is believed that this will only make a difference of one percent or less in which data is identified as being considered as "high-frequency". But as stated above, this study does not make any extreme claims as to what is or what is not a high frequency collocation, but rather simply provides a methodology and the results of that methodology to identify such items, and shows that this methodology produces very good results while it acknowledges that these results are not definite.

Future research

As discussed in the previous section, various limitations in this study leave the door open to a number of future research paths that could and should be taken. It is the hope of this researcher that more research is done that either builds upon what was accomplished in this study, or refutes its findings and proves better and/or more efficient or more useful ways to accomplish the task that was at hand.

First, this study only conducted L1-L2 contrastive analysis with Japanese. In fact, if the most ideal materials could be created, then there would be translations of these contents and L1-L2 contrastive analysis conducted for all learner L1s. This, however, is a tremendous amount of work and so researchers and translators will need to collaborate in the future if this is ever to be achieved.

Also, as discussed in previous sections, a number of technological limitations existed, and this created the need to use software that was not designed for the task at hand. This less than ideal approach led to very complex and time consuming methodologies. Thus, future research should inquire as to ways in which software could be improved upon, such as for dealing with colligational issues. It would be ideal if software existed which could easily and accurately part-of-speech tag certain categories of words which could then be easily counted together.

Furthermore, a number of steps in this study had to be conducted manually. If technological solutions existed, then a significant amount of time could be saved and much more accurate work could be done. Corpus compilers should note the difficulty in using dispersion and chronological data to accurately identify items that could be considered as unbalanced by native speakers and the necessity this study found for manual checking. Concordance software developers could also note the necessity for manual extending of MWUs beyond their cores, and possibly discover automated solutions.

In addition, a tremendous amount of work still needs to be done to determine the best way to actually teach the items identified in this study. The first step (identifying the actual items) has been taken. So, from now, future research should examine the best way to teach such items. Should these items be studied in an isolated way or with full example sentences? Or should they be within larger reading passages? Should learners study more than one MWU connected to a core lemma within it (*political activists, political parties, political leaders*) or should items be rather studied in their frequency rank order? Moreover, what kind of time-frame could these items be mastered in? How long would it take learners to master approximately 11,000 MWUs? The answers to these questions remain to be seen.

Clearly, a tremendous amount of research still needs to be conducted before we can truly improve upon the efficacy of obtaining collocational fluency. It is this researcher's hope that future researchers, materials writers, translators, software developers and ESL practitioners collaborate more to enable the community to answer these difficult questions in the most expedient way possible. Learners need an answer to these questions today, and it is our job and duty as educators to provide them with what they need to help them achieve their learning goals.

Conclusion

Objectives and rationale

This study aimed to identify the collocational exemplars of high-frequency English vocabulary. Researchers have been aware of the value of collocational fluency for years. Language is stored in the brain in such collocational chunks. This makes native speaker-like language processing possible. Without such knowledge, a listener must process each and every word separately, which is a very inefficient process. In addition, the literal processing of words by second language learners can often lead to confusion in that chunks are not always a sum of their parts. Such knowledge also enables a second language speaker to improve upon their language production speed as well, retrieving and producing language chunks rather than stringing words together one by one. Possessing such knowledge also helps non-native speakers avoid errors such as when they may attempt to directly translate from their L1 a phrase that is said in a different way in the L2. Moreover, learning such chunks is actually easier in comparison to learning isolated vocabulary. Each word in a chunk has the potential to serve as a mnemonic hook for the others. The mastering of these chunks not only enables learners to master vocabulary and the way that vocabulary naturally collocates in the L2, but also the L2's grammar implicitly through the formulations. In fact, this is more akin to the way native speakers learn their L1's grammar. For example, if you ask a native speaker to explain the grammar behind their language choices, they will often struggle. However, what they say will be usually be perfectly grammatical. This is because they have, to an extent, intuitively mastered their language's grammar through mastery of such chunks.

This constituted a major gap in the research in that no such resource, other than very small lists (in the hundreds) or very large dictionary-like lists (in the tens of thousands), existed. Because of this lack of a resource, teachers and materials writers could not help learners study such knowledge explicitly. When learning materials do not highlight such items, it is problematic because previous research has shown that a learner's attention needs to be brought to such items for them to learn them. This lack of focus on them has resulted in a severe lack of collocational fluency in English among a variety of learner groups throughout the world.

Approach

From the beginning of this study significant challenges were posed. In fact, just the word 'collocation' itself has been under debate for years and there is still a significant amount of disagreement about what should and what should not be considered a 'collocation'. However, very early on in this study there was the realization that this study was not looking for what previous researchers usually referred to as 'collocations', but rather it was looking for the MWUs most representative of high-frequency, lemmatized concgrams. More generally speaking, what learners need to learn to truly master high-frequency vocabulary and the way they co-occur with each other became the focus, rather than what is or is not a 'collocation'. This way of thinking and its practicality and flexibility was paramount to helping achieve this study's research goals.

At the beginning of this study, there were three paths toward approaching collocations that a decision needed to be made about. Should collocations be approached from a semantic, a structural, or from a lexical approach? For the purposes of this study, the semantic approach was not deemed appropriate because it did not suffice in explaining the idiosyncratic nature of how many formulations collocate. However, in regards to the structural and lexical approaches, it did not end up being an 'either/or' decision. There seemed to be validity in both approaches for the purpose of this study, and thus both were integrated. A lexical approach was taken because it had the highest potential to help achieve the goal of identifying specific items to study explicitly. For such a task, the structural approach left too many questions in regards to which words should/could be places in grammatical matrices, such as with *[verb] a [noun]*. Despite this, there was the realization that there are times in which including a structural approach in addition to the findings that a lexical approach produces is ideal. This is clear when you consider, for example, that *early in the X century* is the MWU most representative of the lemma *early/century* rather than *century earlier*, where *X* can only be identified by counting years that occur when *early* and *century* occur together as a grammatical matrix.

Adopting a lexical approach then led to the need to select a corpus to source data from. The COCA was the most logical choice in that it is freely accessible and aids future replicability of this study. In addition, the COCA is significantly larger and more balanced in comparison to the corpus that much previous research has utilized (the BNC). The COCA also provides lemmatized concgram lists and thus was ideal for this current study. However, analyzing the data it provided was not an easy task. Numerous experiments needed to be conducted to get the

data to the point where it could be practically used to teach the items explicitly. For example, it was necessary to determine the most ideal frequency cut-off for the concgrams and it was also necessary to develop a methodology to find only items that had balanced dispersion and chronological data because the goal was to identify only items that had general value for learners. It was also necessary to develop a custom methodology to help identify any items that would benefit from a colligational treatment, such as *early/century* did above.

The ultimate goal was to create a resource that could be studied or taught explicitly, so at some point the lemmatized concgram data (such as *early/century*) needed to be analyzed to extract the collocational exemplar (the MWU) most representative of how those two lemma usually co-occur. This required a software solution which did not exist. The author of the most downloaded concordance software in the world, Laurence Anthony, was thus recruited to help develop custom software specifically for this study's purpose. Without it, this study goals would not have been accomplished so easily. After much development and consideration, this software was completed, and the MWUs were identified.

After the items were identified, they needed to be translated to make L1/L2 congruency analysis possible. This step made it possible to identify incongruent items, which have a higher learning burden. By identifying them, teachers would then know which items to spend more class time on. To conduct such an analysis, all the MWUs identified needed to be translated into the L1 in question, which was not an easy task since there were over 11,000 of them. However, with the aid of volunteer translators, this step in the research was also completed.

Another step that was necessary was to conduct a semantic transparency analysis because this aspect of the MWUs can also significantly affect their learning burden. The MWUs were thus rated as to their level of semantic transparency from literal to idiomatic.

In addition, since the ultimate goal of this study had the practical needs of learners in mind, it was deemed insufficient to simply provide a list of MWUs and their translations to learners. To fully understand a MWU, it is occasionally necessary to study it within the larger context of a full example sentence. The details such a sentence provides help learners to note the most appropriate way to use the MWU. To achieve this goal, native speakers were recruited to create an example sentence for each of the 11,000 MWUs. They were instructed to be careful not to add additional learning burden to the MWU by avoiding the use of low-frequency

vocabulary in the contexts they created. But the question then arose as to whether or not native speakers could be relied upon to accomplish such a task while only relying on their intuition. Thus, an analysis of the data they produced was conducted as well.

Finally, this study aimed to determine Japanese university students' knowledge of such items, and whether or not item frequency, L1/L2 congruency, or semantic transparency played a role in affecting that knowledge. A balanced selection of items was taken with their criteria in mind, and a group of Japanese university students were tested on their productive knowledge of the items. These results were then analyzed to determine whether or not the aforementioned criteria truly did play a role in affecting their knowledge of them.

Results

Although for many years researchers and ESL practitioners have been aware of the value of explicitly learning such items, a comprehensive resource did not exist. By examining previous research and the steps that would need to be taken to achieve such a goal, it became clear why such a resource was not yet available. A number of obstacles stood in the way of this being accomplished. The vast scale of the work that needed to be done, and the inability of one individual accomplishing it was one of them. The scope of the data that needed to be analyzed and translation required was staggering and necessitated the recruiting of a number of volunteers to devote their time for many years. Another was the lack of established methodologies. This study thus needed to create and experiment with a number of different methods. There were also technological limitations. The type of software needed to complete certain steps in this research simply did not exist. It is also very important to note that this study took a post-positivist approach, and thus the research had to employ measures that approximated reality, while admittedly possessing weaknesses that were unavoidable. With all of these limitations in mind, this study should be considered as having achieved the goals it set forth.

In regards to the frequency cut-off experiment, as already mentioned, measures that approximated reality had to be taken. Just as there will never be any unequivocal frequency cut-off for high-frequency vocabulary, there also will not for lemmatized concgrams. However, vocabulary lists that justify their cut-offs to the best ability possible, such as at 3,000 word families (Nation, 2004), is certainly a practical trade-off, and unavoidable regardless. This study

chose a cut-off which resulted in a list of approximately 11,000 lemmatized concgrams which consisted of approximately 3,000 word families. These word families had high coverage of high-frequency vocabulary and had the vast majority of them fall within what would be considered as high-frequency vocabulary. Thus, this cut-off was deemed as producing very good results with the practical goal of explicitly teaching the items since previous research stated that such high-frequency vocabulary should be taught explicitly. Moreover, it is ideal to teach such vocabulary within the MWUs they most commonly occur in.

Some of the steps this study took to achieve its goals did not prove fruitful though. However, these results are very informative for future researchers to avoid or devise ways to better deal with such issues. An example would be the steps that were taken to identify items that had only balanced dispersion and chronological data. Although the steps taken did identify such items to an extent, they were still quite inaccurate as far as native speaker intuition was concerned. Despite a variety of different parameters being used, the results were either too inclusive or too exclusive. Thus, this method using such a computer data analysis was deemed to be inaccurate and manual checking of items using native speaker intuition was considered to be essential. Such manual checking was conducted, and while time-consuming, it was accomplished. So, in the end the type of data that was aimed for, only items that were considered to occur among a wide variety of texts (balanced dispersion) and items which were not dated, too modern, or only occurring during a specific time period (balanced chronological data distribution), was identified.

Furthermore, the steps this study took to grapple with colligational issues also did not prove totally fruitful. Despite being important and clearly improving upon the results to a small extent, the extreme complexity and significant amount of time required to complete the steps is still currently an issue. No dedicated software or methodology exists to deal with such issues, and thus this study had to resort to using non-dedicated software for the task and to also devise its own methodology. Improvements were clearly made to the data, but only to a very small percentage of it. So, the results show that while such steps should be taken to improve results, a less cumbersome way to achieve such goals would be ideal.

In contrast, the results of the experiment which utilized native speaker intuition to identify MWUs worthy of expanding beyond their most frequent exemplar proved to be

extremely fruitful. A majority of the items examined benefitted from such a treatment and thus such a step should be considered essential. This step did require the manual checking of items by native speakers, which is quite time consuming though. However, such an analysis does not seem to be possible through a technological solution, and thus this issue may simply be unavoidable.

Surprising results include the percentage of MWUs that were considered to have high semantic transparency. The majority of the way high-frequency vocabulary collocate with each other was shown to be in literal formulations. Such items would not be included in what is generally thought of as being collocations. But high-frequency vocabulary needs to be taught, and it is an established fact that teaching isolated vocabulary is not ideal. Such vocabulary should be taught in the MWUs they occur most often in. Thus, to teach high-frequency vocabulary and develop 'collocational' knowledge teaching literal formulations becomes unavoidable.

Results regarding L1-L2 congruency analysis were as expected with a large percentage of the items analyzed being non-congruent to some extent. Previous research indicates that items with low L1-L2 congruency will have a higher learning burden. Thus, by creating a methodology to rate congruency between English and Japanese and conducting an analysis of over 11,000 MWUs, this part of the research produced very useful data in that a large amount of items which will need additional study time because of their higher learning burden have been identified.

Another successful aspect of the study was the results in regards to the reliability of native speaker intuition for selecting high-frequency vocabulary when creating the context to support teaching the MWUs. Native speaker intuition proved extremely useful in that the vast majority of the added content to the MWUs (the example sentences created by them) were high-frequency. Creating such custom content was necessary and serves as further support to help learners understand proper usage of the MWUs. So although it was quite time consuming it is an unavoidable step that must be taken. Yet, before this study was conducted it was unclear or not whether native speakers could be relied upon when posed with the task of creating examples sentences for the MWUs while avoiding the use of low-frequency vocabulary. If they did add low-frequency vocabulary, they would be adding unnecessary learning burden to the task of

mastering the MWUs themselves. However, the results of the data analysis of the content created revealed that the native speakers did not add any significant amount of learning burden. Thus, this step in the process was deemed successful.

The final step in this research project was to test Japanese university students' knowledge of the MWUs identified. The results of this experiment were as expected. Japanese university students have very little to no knowledge of the MWUs. This is understandable because, as this study pointed out many times, a resource that identifies the most common MWUs that high-frequency vocabulary occur in does not exist. Therefore, teachers and materials writers cannot refer to anything to teach them directly. So, it is not surprising that the learners in question had very little knowledge of the items. There were some limitations to the results of this step in the study though. The test did include a balanced sample of items in regards to frequency, L1-L2 congruency, and semantic transparency and test students from a wide range of TOEFL scores to determine if any of these factors play a role in the students' knowledge of the items. However, L1-L2 congruency or semantic transparency was not shown to be a factor in affecting knowledge, which is contrary to what previous research stated and what logically makes sense. TOEFL scores also did not correlate with item knowledge. Although, this was probably due to the fact that the item knowledge was so low across the board that the results did not even register the effects of these criteria.

Final thoughts

It is the firm belief of this researcher that this study achieved the task it set forth to with good results. There are clear limitations to the results and how they can be interpreted, but this study still filled major gaps in the research. First, it created a resource that can now be practically used by a specific group of learners. This resource can also be expanded upon by future researchers for different groups of learners as well with L1 translation and congruency analysis. Furthermore, a number of new methodologies were created to achieve the tasks in this study, and these can now be used or improved upon by other researchers in the future as well.

New discoveries were made as well that can inform and improve upon corpus creation and corpus data access. Discoveries revealed limits to what can be achieved using corpus data, such as were found with dispersion and chronological data. There were even revelations that

contradict previous thoughts about co-occurring words, such as how the vast majority of items identified were actually literal formulations. Furthermore, certain aspects of this study showed how native speaker intuition can be relied upon to a large extent for high-frequency vocabulary usage in context creation.

This study also led to the development of new concordance software and the highlighting of how there is still a lack of dedicated software that can accomplish the task of dealing with colligational isssues. In addition, the importance of conducting L1-L2 congruency analysis to identify items that have a higher learning burden was confirmed. A severe lack of collocational fluency in Japanese university students was also revealed.

It is the hope of this researcher that much more research is conducted on this topic. Further L1-L2 congruency analysis with other L1s is certainly called for. Furthermore, improvements to the methodologies that this study utilized are expected as well, and in turn, the development of new and improved resources. This study proved extremely time-consuming and required a lot of man-power, and thus it is hoped that some technological solutions can eventually be found that would help automate some of the steps that were taken. From this point forward, it would be interesting to see the results of actually explicitly teaching the items identified in this study. Many questions still remain, such as how much time it would take to master them, what study methods are the most efficient, and resulting improvements on standardized tests, if any. It would also be interesting to see how the resource that this study created could be used as a reference for materials creation. Clearly, much more research needs to be done on this topic and it is the sincere wish of this researcher that practitioners in the field make an effort to collaborate and accomplish these tasks because by enhancing learners' collocational knowledge through a variety of integrated exercises which focus on the ultimate goal of daily communication with English speakers, learners' listening, speaking, reading and writing fluency has a higher potential for improving. Since our ultimate goal as ESL teachers and researchers is to help learners achieve fluency in the most efficient way possible, the resource that this study resulted in should be considered as a tool with the potential to help achieve this.

References

Abdul-Fattah, H. (2001). Collocation: A missing chain from Jordanian basic education stage English language curriculum and pedagogy. *Dirasat, Humand and Social Sciences, 28* (2), 582-596.

Ackermann, K., & Chen, Y. (2013). Developing the academic collocation list (ACL): A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes, 12*(4), 235-247.

Aghbar, A. (1990, October). *Fixed expressions in written texts: Implications for assessing writing sophistication.* Paper presented at the Meeting of the English Association of Pennsylvania State System Universities, Pennsylvania, U.S.A.

Al-Zahrani, M. (1998). *Knowledge of English lexical collocations among male Saudi college students majoring in English at a Saudi university* (Unpublished doctoral dissertation). University of Pennsylvania, Pennsylvania, U.S.A.

Alexander, R. (1984). Fixed expressions in English: Reference books and the teacher. *ELT Journal, 38*(2), 127-134.

Almela, M., & Sanchez, A. (2007). Words as "lexical units" in learning/teaching vocabulary. *International Journal of English Studies, 7*(2), 21-40.

Anderson, R., & Nagy, W. (1991). Word meanings. In R. Barr, M. Kamil, P. Mosenthal, and P.D. Pearson (Eds.), *Handbook of Reading Research (Vol. 2)* (pp. 690-724). New York: Longman.

Anthony, L. (2011). *AntConc* (*Version 3.2.2*) [Computer Software]. Tokyo, Japan: Waseda University, Retrieved from http://www.antlab.sci.waseda.ac.jp/

Anthony, L. (2013). *AntWordPairs* (*Version 1.0.2*) [Computer Software]. Tokyo, Japan: Waseda University, Available on request.

Arabski, J. (1979). *Errors as indicators of the development of interlanguage*. Katowice: Uniwersytet Slaski.

Avery, P., & Baker, J. (1997). Mapping learning at the secondary level. *Clearing House, 70*(5), 279-285.

Bahns, J., & Eldaw, M. (1993). Should we teach ESL students collocations? *System, 21*(1), 101-114.

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography, 6*(4), 253-279.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371-405.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

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

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python.* Boston, O'Reilly Media.

Biskup, D. (1992). L1 influence on learners' renderings of English collocations: A Polish/German empirical study. In P. Arnaud and H. Bejoint (Eds.), *Vocabulary and Applied Linguistics* (pp. 1-12). London: Macmillan.

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

Bogaards, P. (2001). Lexical units and the learning of foreign language vocabulary. *Studies in Second Language Acquisition, 23*, 321-343.

Bolinger, D. (1968). *Aspects of language*. New York: Harcourt Brace Jovanovich, Inc.

Bolinger, D. (1976). Meaning and memory. *Forum Linguisticum, 1(1)*, 1-14.

Bonk, W. (2000). *Testing ESL learners' knowledge of collocations*. Illinois: U.S. State Department.

Burston, J. (2007). CALICO software review: WordChamp. *CALICO Journal, 24*(2), 473-486.

Channell, J. (1981). Applying semantic theory to vocabulary teaching. *English Language Teaching Journal, 35*, 115-122.

Chan, T., & Liou, H. (2005). Online verb-noun collocation instruction with the support of a bilingual concordancer. Selected papers from the fourteenth international symposium on English teaching (pp. 270 – 281). Taipei: Crane Publishing Co., Ltd.

Chen, P. (2002). *A corpus-based study of the collocational errors in the writings of the EFL learners in Taiwan* (Unpublished master's thesis). National Taiwan Normal University, Taipei, Taiwan.

Cheng, W., Greaves, C., & Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics, 11*(4), 411-433.

Chomsky, N. (1957). *Syntactic structures.* The Hague: Mouton and Company.

Chon, Y., & Shin, D. (2009). Collocations in L2 writing and rater's perceived writing proficiency. *Korean Journal of Applied Linguistics, 25*(1), 101-129.

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

Cobb, T. (2013). *Vocabprofile*. Retrieved from http://www.lextutor.ca/vp/

Conzett, J. (2000). Integrating collocation into a reading and writing course. In M. Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 70-86). Hove, England: Language Teaching Publications.

Cortes, V. (2002). Lexical bundles in freshman composition. In R. Reppen, S.M. Fitzmaurice & D. Biber (Eds.), *Using corpora to explore linguistics variation* (pp. 131-145). Amsterdam: John Benjamins Publishing Company.

Cowan, L. (1989). *Towards a definition of collocation* (Unpublished master's thesis). Concordia University, Montreal, Quebec, Canada.

Cowie, A. (1978). The place of illustrative material and collocations in the design of a learner's dictionary. In P. Strevens (Ed.), *In Honour of A.S. Hornby* (pp. 127-139). Oxford: Oxford University Press.

Cowie, A. (Ed.) (1998). *Phraseology: theory, analysis, and applications.* Oxford: Oxford University Press.

Daulton, F. (2008). *Japan's built in lexicon of English-based loanwords.* Clevendon: Multilingual Matters Ltd.

Davies, M. (2008). *The Corpus of Contemporary American English: 425 million words, 1990-Present*. Retrieved from at http://corpus.byu.edu/coca/

Davies, M. (2010). *Word List Plus Collocates*. Retrieved from http://www.wordfrequency.info/purchase1.asp?i=c5a

Davies, W. (2013). *How to study psychology.* London: Psychology Press.

Dechert, H. (1983). How a story is done in a second language. In C. Faerch, and G. Kasper, (Eds.), *Strategies in Interlanguage Communication* (pp.175-195). London: Longman.

Dechert, H., & Lennon, P. (1989). Collocational blends of advanced second language learners: A preliminary analysis. In W. Olesky (Ed.), *Contrastive pragmatic* (pp. 131-168). Amsterdam: Benjamins.

DeCock, S., Granger, S., Leech, G. and McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learner English on computer* (pp.67-79). London and New York: Longman.

Digital Volcano (2011). *Textcrawler (Version 2.5)* [Computer Software] Retrieved from http://www.digitalvolcano.co.uk/content/textcrawler

Doughty, C., & Williams, J. (1998). Pedagogical choices in focus on form. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 197-262). New York: CUP.

Durrant, P. (2014). Corpus frequency and second language learners' knowledge of collocations: A meta-analysis. *International Journal of Corpus Linguistics, 19*(4), 443-477.

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

Ellis, N. (1997). Vocabulary acquisition: word structure, collocation, word-class, and meaning. In N. Schmitt and M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 122–139). Cambridge: Cambridge University Press.

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

Ellis, N., Simpson-Vlach, R., & Maynard, C. (2008). 'Formulaic language in native and second-language speakers: psycholinguistics, corpus linguistics, and TESOL,' *TESOL Quarterly 42*(3), 75–96.

Ellis, R. (1985). *Understanding second language acquisition.* Oxford: Oxford University Press.

Ellis, R. (1994). *The study of second language acquisition.* Oxford: Oxford University Press.

Farghal, M., & Obeidat, H. (1995). Collocations: A neglected variable in EFL. *IRAL, 33*(4), 315-331.

Fayez-Hussein, R. (1990) Collocations: The missing link in vocabulary acquisition amongst EFL learners. In J. Fisiak (Ed.), *Papers and studies in contrastive linguistic: The Polish English contrastive project, 26* (pp. 123-126). Poznan, Poland: Adam Mickiewicz University.

Firth, J. (1957). A synopsis of linguistic theory. 1930-1955. In *Studies in linguistic analysis* (pp. 1-32), reprinted in F. Palmer (Ed.), *Selected papers of J.R. Firth 1952-59* (pp. 168-205). London: Longman.

Francis, G. (1993). A corpus-driven approach to grammar: Principles, methods and examples. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 137-156). Amsterdam: John Benjamins.

Furukawa, J., Ford, B., Ayson, E., Cambra, K., Takahashi, L. & Yoshina, K. (1998, January). *Effects of a cognitive processing strategy on spelling, definitions, and reading*. Paper presented at the annual meeting of the Hawaii Educational Research Association. Honolulu, Hawaii, U.S.A.

Gairns, R., & Redman, S. (1986). *Working with words. A guide to teaching and learning vocabulary.* Cambridge: CUP.

Gasser, M. (1990). Connectionism and universals of second language acquisition. *Studies in Second Language Acquisition, 12,* 179-199.

Gitsaki, C. (1996). *The development of ESL collocational knowledge* (Unpublished doctoral dissertation). University of Queensland, Brisbane, Australia.

Goto, K. (2005). *GoTagger (Version 0.7)* [Computer Software]. Retrieved from http://web4u.setsunan.ac.jp/Website/GoTagger.htm#

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.), *Phraseology, Theory, Analysis and Applications* (pp. 145-160). Oxford: Calrendon.

Grant, L., & Bauer, L. (2004). Criteria for re-defining idioms: Are we barking up the wrong tree? *Applied Linguistics 25*(1), 38-61.

Gries, S. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics 13*(4), 403-437.

Grucza, B., & Jaruzelska, H. (1978). Typowe bledy popelnanie przez kandydatow jezyku niemieckim: bledy gramatyczne i leksykalne. In F. Grucza (Ed.), *Z Problematyki Bledow Obcojezycznych*, (pp. 124-143). Warsaw, Poland: WSP.

Guerra, A. (2015). *Linear transformations of semantic spaces for word-sense discrimination and collocation compositionality grading*. (Unpublished doctoral dissertation). The University of Dublin, Dublin, Ireland.

Gyllstad, H. (2005). Words that go together well: developing test formats for measuring learner knowledge of English collocations. *International Journal of English Studies 7*(2), 127-157.

Gyllstand, H. (2007). *Testing English collocations: Developing receptive tests for use with advanced Swedish learners* (Unpublished doctoral dissertation). Lund University, Lund, Sweden.

Hadley, A. (2001). *Teaching language in context.* Boston: Heinle and Heinle.

Halliday, M., & Sinclair, J. (1966). Beginning the study of lexis. In C. Bazell, J. Catford, M. Halliday & R. Robins (Eds.), *In memory of J.R. Firth* (pp. 410-430). London: Longman.

Hausmann, F. (1984). Wortschatzlernen ist Kollokationslernen. Zum Lshren und Lernen franzosischer Wortverbindungen. *Praxis des neusprachlichen Unterrichts, 31,* 395-406.

Heatley, A., Nation, P., & Coxhead, A. (2002). RANGE program. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation

Hill, J. (2000). Revising priorities: from grammatical failure to collocational success. In M. Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 47-67). Hove, England: Language Teaching Publications.

Hill, J., Lewis, M., & Lewis, M. (2000). Classroom strategies, activities, and exercises. In M. Lewis (Ed), *Teaching Collocation: Further developments in the lexical approach* (pp. 88-116). Hove, England: Language Teaching Publications.

Hoey, M. (1991). *Patterns of lexis in text.* Oxford: Oxford University Press.

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

Hopkins, G., & Bean, T. (1999). Vocabulary learning with the verbal-visual word association strategy in a native American community. *Journal of Adolescent and Adult Literacy, 42,* 274-281.

Howarth, P. (1996). *Phraseology in English academic writing. Some implications for language learning and dictionary making.* Tübingen: Niemeyer.

Hsu, J. (2002). *Development of collocational proficiency in a workshop on English for general business purposes for Taiwanese college students* (Unpublished doctoral dissertation). Indiana University of Pennsylvania, Pennsylvania, U.S.A.

Hsu, J., & Chiu, C. (2008). Lexical collocations and their relations to speaking proficiency of college EFL learners in Taiwan. *The Asian EFL Journal Quarterly, 10*(1),181-204.

Hsu, L. (2005). *The effect of lexical collocation instruction on Taiwanese college EFL learners' listening comprehension* (Unpublished master's thesis). National Kaohsiung First University of Science and Technology, Taiwan.

Huang, L. (2001). Knowledge of English collocations: An analysis of Taiwanese EFL learners. In C. Luke and B. Rubrecht (Eds.). *Texas Papers in Foreign Language Acquisition: Selected Proceedings from the Texas Foreign Language Education Conference* (pp. 113-132). Austin, Texas: Texas University.

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

Hwang, K., & Nation, P. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System* 23(1), 35-41.

Hyland, K. (1994). The learning style of Japanese students. *JALT Journal, 16*(1), 55-74.

Jaen, M. (2007). A corpus-driven design of a test for assessing the ESL collocational competence of university students. *International Journal of English Studies, 7*(2), 127-147.

Japanese Ministry of Justice (n.d.). Foreign national residents by nationality. Retrieved from http://www.stat.go.jp/data/nenkan/zuhyou/y0214000.xls

Jiang, N., & Nekrasova, T. (2007). The processing of formulaic sequences by second language speakers. *Modern Language Journal, 91*(3), 357-377.

Jones, S., & Sinclair, J. (1974). English lexical collocations: a study in computational linguistics. *Catiers de Lexicologie, 23*(2), 15-61.

Kaneko, E. (2008). *An analysis of oral performance by Japanese learners of English* (Unpublished doctoral dissertation). The University of Wisconsin, Wisconsin, U.S.A.

Kennedy, G. (1990). Collocations: Where grammar and vocabulary teaching meet. In S. Anivan (Ed.), *Language teaching methodology for the nineties* (pp. 215-229). Singapore: RELC.

Keshavarz, M., & Salimi, H. (2007). Collocational competence and cloze test performance: A study of Iranian EFL learners. *International Journal of Applied Linguistics, 17*(1), 81-92.

Kilgarriff, A., Atkins, S. & Rundell, M. (2007, July). *BNC design model past its sell-by*. Paper presented at the Corpus Linguistics Conference. Birmingham, UK.

Kjellmer, G. (1984). Some thoughts on collocational distinctiveness. In J. Aarts and W. Meijs (Eds.), *Computer corpora in English language research* (pp. 163-171). Bergen: Norwegian Computing Centre for the Humanities.

Kjellmer, G. (1987). Aspects of English collocations. In W. Meijs (Ed), *Corpus linguistics and beyond* (pp. 133-140). Amsterdam: Rodopi.

Kjellmer, G. (1990). Patterns of collocability. In J. Aarts and W. Meijs (Eds.) *Theory and Practice in Corpus Linguistics* (pp. 163-178). Amsterdam: Rodopi.

Kjellmer, G. (1994). *A dictionary of English collocations: Based on the Brown corpus*. Oxford: Claredon Press.

Korosadowicz-Struzynska, M. (1980). Word collocations in FL vocabulary instruction. *Studia Anglica Posnaniensia, 12,* 109-120.

Koya, T. (2004). Collocation research based on corpora collected from secondary school textbooks in Japan and in the UK. *Dialogue, 3,* 7-18.

Kuiper, K. (1996). *Smooth talkers.* Hillsdale, NJ: Lawrence Erlbaum.

Larsen-Freeman, D., & Long, M. (1991). *An introduction to second language acquisition research.* White Plains, NY: Longman.

Laufer, B. (1988, April). *Ease and difficulty in vocabulary learning: some teaching implications*. Paper presented at the 22nd Annual Meeting of the International Association of Teachers of English as a Foreign Language. Edinburgh, Scotland.

Laufer, B. (1990). Ease and difficulty in vocabulary learning: Some teaching implications. *Foreign Language Annals, 23*(2), 147-155.

Laufer, B., & Eliasson, S. (1993). What causes avoidance in L2 learning: L1-L2 difference, L1-L2 similarity, or L2 complexity? *Studies in second language acquisition, 15*(1), 35-48.

Laufer, B., & Girsai, N. (2008). Form-focused instruction in second language vocabulary

learning: A case for contrastive analysis and translation. *Applied Linguistics, 29*(4), 694-716.

Lennon, P. (1996). Getting "easy" verbs wrong at the advanced level. *International Review of Applied Linguistics, 34,* 23-36.

Lesniewska, J., & Witalisz, E. (2007). Cross-linguistic influences on L2 and L1 collocations. *EUROSLA Yearbook, 7,* 27-48.

Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward.* Hove: Language Teaching Publications.

Lewis, M. (2000). Language in the lexical approach. In M. Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 8-10). Hove, England: Language Teaching Publications.

Lewis, M. (2000). There is nothing as practical as a good theory. In M. Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 10-27). Hove, England: Language Teaching Publications.

Lien, H. (2003). *The effect of collocation instruction on the reading comprehension of Taiwanese college students* (Unpublished doctoral dissertation). Indiana University of Pennsylvania, Pennsylvania.

Lin, Y. (2002). *The effects of collocation instruction on English vocabulary development of junior high school students in Taiwan* (Unpublished master's thesis). National Kaohsiung Normal University, Kaohsiung, Taiwan.

Lin, Y. (2004). The effects of collocation instruction on vocabulary development. In W. Dai, Y. Leung, P. Chen and K. Cheung (Eds.), *Selected Papers from the Thirteenth International Symposium on English Teaching and Learning*, (pp. 188–197). Taipei: Crane Publishing Co.

Lin, W., Hsiao-Ching, Y., & Ho-Ping, F. (2003). English vocabulary knowledge of first-year university students: Vocabulary size and collocational knowledge. In *The Proceedings of the 2003 Conference and Workshop on FEFL and Applied Linguistics* (pp. 202-213). Taipei: Crane Publishing Co.

Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English.* Malamo, Sweden: Liber Forlag Malmo.

Liu, C. (1999). A study of Chinese Culture University's freshmen's collocational competence: "Knowledge" as an example. *Hwa Kang Journal of English Language and Literature, 5,* 81-99.

Liu, L. (2002). *A corpus-based lexical semantic investigation of verb-noun miscollocations in Taiwan learner's English* (Unpublished master's thesis). Tamkang University, Taipei, Taiwan.

Liu, E., & Shaw, P. (2001). Investigating learner vocabulary: A possible approach to looking at EFL/ESL learners' qualitative knowledge of the word. *International Review of Applied Linguistics, 39*, 171-194.

Lorenz, G. (1999). *Adjective intensification – learners versus native speakers: A corpus study of argumentative writing*. Amsterdam: Rodopi.

van der Meer, A. (1998). Collocations as one particular type of conventional word combinations, their definition and character. *EURALEX '98, 1,* 313-322.

Mackin, R. (1978). On collocations: Words shall be known by the company they keep. In P. Strevens (Ed.), *In honour of A.S. Hornby* (pp. 149-165). Oxford: Oxford University Press.

Martinez, A. (2011). Collocational analysis of a sample corpus using some statistical measures: An empirical approach. *Escuela Oficial de Idiomas, 25,* 763-768.

Martinez, R., & Schmitt, N. (2012). A phrasal expression list. *Applied Linguistics, 33*(3), 299-320.

Marton, W. (1977). Foreign vocabulary learning as problem no. 1 of language teaching at the advanced level. *Interlanguage Studies Bulletin, 2*(1), 33-57.

Meijs, W. (1992). Inferences and lexical relations. In G. Leitner (Ed.), *New directions in English language corpora: Methodology, results, software developments* (pp. 123-152). Berlin: Mouton de Gruyter.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97.

Mitchell, T. (1971). Linguistic 'goings on': Collocations and the other lexical matters arising on the syntagmatic record. *Archivum Linguisticum, 2,* 35-69.

Monbusho (2003). *The government guidelines for teaching in Japan.* Retrieved from www.mext.go.jp/b_m enu/houdou/11/03/990302a/990302p.htm

Moon, R. (1994). The analysis of fixed expressions in text. In M. Coulthard (Ed.), *Advances in Written Text Analysis,* (pp. 117-135). London: Routledge.

Moon, R. (1997). Vocabulary Connections: Multi-Words Items in English. In N. Schmitt and M. McMarthy, M. (Eds.). *Vocabulary. Description, Acquisition and Pedagogy* (pp.40-63). Cambridge, CUP.

Murphy, J. (1983, March). Words: What goes with what? *Paper presented at the 17th Annual Convention of Teachers of English to Speakers of Other Languages*. Toronto, Ontario, Canada.

Myers, J., & Chang, S. (2009). A multiple-strategy-based approach to word and collocation acquisition. *International Review of Applied Linguistics, 47,* 179-207.

Nagy, W., Herman, P., & Anderson, R. (1985). Learning words from context. *Reading Research Quarterly*, *20,* 233-253.

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

Nation, P. (2001b). How many high frequency words are there in English? In M. Gill, A.W. Johnson, L.M. Koski, R.D. Sell and B. Wårvik (Eds.), *Language, Learning and Literature: Studies Presented to Håkan Ringbom* (pp. 167-181). English Department Publications 4, Åbo Akademi University, Åbo.

Nation, P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a Second Language: Selection, Acquisition, and Testing* (pp. 3-13). Amsterdam: John Benjamins.

Nation, P. (2008). *Teaching Vocabulary: Strategies and Techniques.* Boston: Heinle.

Nation, P., & Meara, P. (2002). Vocabulary. In N. Schmitt (Ed.) *An introduction to applied linguistics* (pp. 35-54). London: Edward Arnold.

Nation, P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle.

Nattinger, J., & DeCarrico, J. (1992). *Lexical phrases and language learning.* Oxford: Oxford University Press.

Nelson, F., & Kucera, H. (1979). *The Brown corpus: A standard corpus of present-day edited American English.* Providence, RI: Department of Linguistics, Brown University.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics, 24*(2), 233-242.

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

Nesselhauf, N., & Tschichold, C. (2002). Collocations in CALL: An investigation of vocabulary-building software in EFL. *Computer Assisted Language Learning, 15*(3), 251-279.

Noel, J. (1992). Collocation and bilingual text. In G. Leitner (Ed.), *New directions in English language corpora: Methodology, results, software developments* (pp. 345-357). Berlin: Mouton de Gruyter.

Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 191-226). London: Longman.

Read, (2001). *Assessing vocabulary.* Cambridge: Cambridge University Press.

Read, (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies, 7*(2), 105-125.

Renouf, A., & Sinclair, J. (1991). Collocational frameworks in English. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics* (pp. 128-143). Harlow: Longman.

Ringbom, H. (1998). Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In S. Granger (Ed.) *Learner English on computer* (pp. 41-52). London and New York: Longman.

Robins, R. (1967). *A short history of linguistics.* London: Longman.

Rogers, J. (2013). How many high-frequency words of English do Japanese university freshmen 'know'? *Kansai Gaikokugo University Journal of Inquiry and Research, 97,* 237-252.

Rodriguez, M., & Sadoski, M. (2000). Effects of rote, context, keyword, and context/keyword methods on retention of vocabulary in EFL classrooms. *Language Learning 50*(2), 385-412.

Rudzka, B., Channell, J., Putseys, Y., & Ostyn, P. (1981). *The word you need: Teacher's book.* London: Macmillan.

Saville-Troike, M. (1984). What really matters in second language learning for academic achievement? *TESOL Quarterly, 18*(2), 199-217.

Schmitt, N. (1997). Vocabulary learning strategies. In N. Schmitt and M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 1–46). Cambridge: Cambridge University Press.

Schmitt, N., and McCarthy, M. (Eds.) (1997). *Vocabulary: Description, Acquisition, and Pedagogy*. Cambridge University Press.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition, 19,* 17-36.

Shin, D. (2006). *A collocation inventory for beginners*. (Unpublished doctoral dissertation). Victoria University of Wellington, Wellington, New Zealand.

Shin, D. (2007). The high frequency collocations of spoken and written English. *English Teaching 6*(1), 199-218.

Simpson-Vlach, R., & Ellis, N. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*(4), 487-512.

Simpson, R., & Mendis, D. (2003). A corpus-based study of idioms in academic speech. *TESOL Quarterly, 37*(3), 419-441.

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

Sinclair, J. (1998). The lexical item. In E. Weigand (Ed.), *Contrastive Lexical Semantics.* Amsterdam: John Benjamins

Skinner, B. (1957). *Verbal behavior.* New York: Appleton-Century-Crofts.

Sokmen, A. (1997). Current trends in teaching second language vocabulary. In N. Schmitt and M. McCarthy (Eds.), *Vocabulary: description, acquisition and pedagogy.* Cambridge: Cambridge University Press.

Someya, Y. (1998). *E-lemma list.* Retrieved from http://www.antlab.sci.waseda.ac.jp/software/resources/e_lemma.zip

Taylor, C. (1983). Vocabulary for education in English. *World Language English, 2*(2), 100-104.

Tseng, F. (2002). *A study of the effects of collocation instruction on the collocational competence of senior high school students in Taiwan* (Unpublished master's thesis). National Taiwan Normal University, Taipei, Taiwan.

Twaddell, F. (1973). Vocabulary expansion in the TESOL classroom. *TESOL Quarterly, 7*(1), 61-78.

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

University Centre for Computer Corpus Research on Language. (n.d.) CLAWS7 tag set. Retrieved from http://ucrel.lancs.ac.uk/claws7tags.html

Wang, C. (2001). *A study of the English collocaiontal competence of English majors in Taiwan* (Unpublished master's thesis). Fu Jen Catholic University, Taipei, Taiwan.

Waring, R., & Takaki, M. (2008). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language, 15,* 130-163.

Webb, S., & Kagimoto, E. (2011). Learning collocations: Do the number of collocates, position of the node word, and synonymy affect learning? *Applied Linguistics, 32*(3), 259-276.

Webb, S., & Nation, P. (2008). Evaluating the vocabulary load of written text. *TESOLANZ Journal, 16,* 1-10.

Wilkins, D. (1972). *The learner's vocabulary for text analysis.* Dortmund: Lensing Verlag.

Wilks, Y. (2005). REVEAL: the notion of anomalous texts in a very large corpus. Tuscan Word Centre International Workshop. Certosa di Pontignano, Tuscany, Italy, 31 June–3 July 2005.

Wood, D. (2004). An empirical investigation into the facilitating role of automatized lexical phrases in second language fluency development. *Journal of Language and Learning, 2*(1), 27-50.

Woolard, G. (2000). Collocation – encouraging learner independence. In M. Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 28-46). Hove, England: Language Teaching Publications.

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

Yorio, C. (1980). Conventionalized language forms and the development of communicative competence. *TESOL Quarterly, 14*(4), 433-442.

Zhang, X. (1993). *English collocations and their effect on the writing of native and non-native college freshmen* (Unpublished doctoral dissertation). Indiana University of Pennsylvania, Pennsylvania, U.S.A.

Zughoul, M. (1991). Lexical choice: Towards writing problematic word lists. *International Review of Applied Linguistics, 29,* 45-58.

List of Appendices

Appendix 23. 3,414 MWUs with an L1-L2 congruency rating of 6 or less

Appendix 24. Example sentences created by native speakers for all 11,208 MWUs

Appendix 25. 50 question collocational fluency test and relevant data

Appendix 26. Collocational fluency test results

List of Tables

[8] Excluding proper nouns such as names and states

List of Figures

List of Related Publications and Resources

Rogers, J. (in press). Teaching collocations. In J. Liontas (Ed.), *The TESOL encyclopedia of English language teaching.* Hoboken, NJ: Wiley-Blackwell.

Rogers, J., Brizzard, C., Daulton, F., Florescu, C., MacLean, I., Mimura, K., … Shimada, Y. (2016). 英語マスター3 千 [English Master 3,000] (Version 1.0) [Mobile application software]. Retrieved from http://itunes.apple.com

Rogers, J., & Reid, G. (2016). What is Japanese university students' knowledge of multi-word units most representative of high-frequency lemmatized concgrams. *The 2nd IRI Research Forum,* 104-113.

Rogers, J., & Murray, B. (2016). On the role of semantic transparency in identifying high-frequency collocations. *The Kyoto JALT Review,* 3, 1-8.

Rogers, J., Brizzard, C., Daulton, F., Florescu, C., MacLean, I., Mimura, K., … Shimada, Y. (2015). 英語マスター1 万 [English Master 10,000] (Version 1.0) [Mobile application software]. Retrieved from http://itunes.apple.com

Rogers, J., Daulton, F., Maclean, I., & Reid, G. (2015). Is native speaker intuition reliable in regards to high-frequency vocabulary usage in context creation? *Kansai Gaikokugo University Journal of Inquiry and Research,* 57-69.

Rogers, J., Brizzard, C., Daulton, F., Florescu, C., MacLean, I., Mimura, K., … Shimada, Y. (2015). On using corpus frequency, dispersion, and chronological data to help identify useful collocations. *Vocabulary Learning and Instruction, 4*(2), 21-37.

Rogers, J., Brizzard, C., Daulton, F., Florescu, C., MacLean, I., Mimura, K., … Shimada, Y. (2014). A methodology for identification of the formulaic language most representative of high-frequency collocations. *Vocabulary Learning and Instruction, 3*(1), 51-65.

Rogers, J., (2013). Corpora: A discussion of corpus resources and their benefits and limitations in regards to informing pedagogy. *Higher Education Research, 3,* 100-104.

Rogers, J. (2013). On how to identify useful collocations and the multi-word units they occur in. *The Kyoto JALT Review,* 1, 64-93.

Appendix 1. Top 5000 lemma of the COCA
(This is just a sample. Full data available upon request)

LEMMA	PART OF SPEECH	FREQUENCY
the	a	22038615
be	v	12545825
and	c	10741073
of	i	10343885
a	a	10144200
in	i	6996437
to	t	6332195
have	v	4303955
I	p	3978265
it	p	3872477
to	i	3856916
that	c	3430996
for	i	3281454
you	p	3081151
he	p	2909254
with	i	2683014
do	v	2573587
on	i	2485306
say	v	1915138
this	d	1885366
they	p	1865580
we	p	1820935
his	a	1801708
but	c	1776767
at	i	1767638
that	d	1712406
not	x	1638830
from	i	1635914
n't	x	1619007
by	i	1490548
she	p	1484869
or	c	1379320
as	c	1296879
what	d	1181023
go	v	1151045
their	a	1083029
can	v	1022775
who	p	1018283
get	v	992596
her	a	969591
if	c	933542
would	v	925515
my	a	919821
know	v	892535
all	d	892102

about	i	874406
make	v	857168
as	i	829018
will	v	824568
up	r	795534
brave	j	5061
dense	j	5061
twist	n	5060
flying	j	5056
devastating	j	5055
devil	n	5051
technician	n	5048
skilled	j	5047
honestly	r	5042
regain	v	5041
manual	n	5040
delight	n	5038
beyond	r	5036
depart	v	5035
severely	r	5035
butterfly	n	5031
vacuum	n	5028
contemplate	v	5027
middle-class	j	5025
low	r	5024
meantime	n	5022
warehouse	n	5020
biography	n	5015
weave	v	5015
speculate	v	5012
organized	j	5011
epidemic	n	5010
seldom	r	5009
photograph	v	5006
plea	n	5000
compelling	j	4995
cargo	n	4991
troubled	j	4991
disturbing	j	4990
naval	j	4990
accusation	n	4987
overwhelm	v	4976
apology	n	4972
convenience	n	4972
copy	v	4970
sometime	r	4938
dictate	v	4935
frustrate	v	4933
accelerate	v	4923
boring	j	4922

praise	n	4896
public	r	4877
fatal	j	4875

Appendix 2. 25,969 lemma pairs resulting from cut-off of one occurrence per million tokens
(This is just a sample. Full data available upon request)

PIVOT WORD	PART OF SPEECH	COLLOCATE	PART OF SPEECH	M.I.	FREQUENCY
will	v	have	v	1.2	157481
do	v	know	v	1.53	124843
do	v	think	v	1.32	93945
year	n	ago	r	5.4	78019
ago	r	year	n	6.71	78015
do	v	want	v	1.36	64614
all	r	right	r	6.38	56425
right	r	all	r	6.35	56420
will	v	say	v	1.09	56336
as	r	well	r	4.61	54302
do	v	how	r	1.02	53162
know	v	how	r	2.47	50339
how	r	know	v	2.48	50321
high	j	school	n	5.06	49398
school	n	high	j	5.06	49390
up	r	pick	v	5.35	46894
already	r	have	v	2.1	46496
back	r	come	v	3.39	46179
come	v	back	r	3.41	46159
up	r	come	v	2.27	45949
see	v	can	v	1.06	43244
right	r	now	r	4.17	42955
now	r	right	r	4.21	42904
can	v	see	v	1.65	42626
go	v	back	r	2.35	40622
can	v	get	v	1.08	40482
go	v	out	r	1.37	38293
will	v	think	v	1.81	37402
would	v	think	v	1.39	35661
do	v	why	r	1.62	35328
how	r	can	v	1.06	35089
out	r	there	r	3.03	34968
there	r	out	r	3.03	34968
no	r	long	r	7.88	34660
come	v	out	r	2.1	34566
out	r	come	v	2.09	34504
up	r	get	v	1.2	34322
find	v	out	r	2.75	34282
out	r	find	v	2.76	34212
can	v	how	r	1.71	34156
could	v	see	v	1.99	32683
get	v	out	r	1.32	31899
out	r	get	v	1.32	31881
sure	j	make	v	4.06	31557
make	v	sure	j	4.08	31533

so	r	far	r	3.74	28739
woman	n	man	n	3.43	28738
man	n	woman	n	3.42	28732
up	r	grow	v	4.09	28269
will	v	know	v	1.21	28220
white	j	folk	n	2.86	500
civil	j	disobedience	n	9.76	500
fresh	j	ginger	n	7.55	500
fresh	j	tomato	n	5.69	500
alternative	j	medicine	n	6.25	500
electrical	j	power	n	4.63	500
school	n	enrollment	n	4.38	500
program	n	component	n	2.58	500
house	n	beach	n	2.39	500
education	n	approach	n	2.1	500
process	n	thought	n	2.27	500
market	n	firm	n	2.51	500
field	n	expert	n	2.76	500
drug	n	benefit	n	2.63	500
project	n	pilot	n	4.11	500
practice	n	session	n	4.65	500
doctor	n	visit	n	3.92	500
plant	n	seed	n	4.45	500
board	n	message	n	3.24	500
author	n	article	n	3.03	500
sound	n	wave	n	4.12	500
article	n	author	n	3.04	500
message	n	board	n	3.22	500
bill	n	energy	n	3.1	500
fish	n	oil	n	3.11	500
bag	n	duffel	n	9.85	500
opinion	n	majority	n	4.35	500
beach	n	house	n	2.56	500
get	v	feel	n	1.96	500
draw	v	gun	n	3.08	500
fly	v	airplane	n	6.2	500
lift	v	eye	n	2.43	500
oppose	v	group	n	2.62	500
assign	v	student	n	3.03	500
ease	v	pain	n	6.1	500
thing	n	through	r	1.05	500
very	r	rapidly	r	2.33	500
am	r	around	r	2.65	500
slightly	r	less	r	3.55	500
rapidly	r	very	r	2.34	500
look	v	alike	r	3.2	500
accept	v	widely	r	5.01	500
business	n	operate	v	2.51	500
space	n	occupy	v	4.87	500
list	n	compile	v	6.83	500

off	r	tip	v	4.13	500
home	r	fly	v	2.68	500
widely	r	accept	v	5.03	500
would	v	satisfy	v	1.45	500
spend	v	study	v	2.35	500
steal	v	try	v	2.14	500

Appendix 3. Cut-off of two occurrences per million tokens resulting in 1,874 word families plus off-list types

Word List	Families (%)	Types (%)	Tokens (%)	Cumulative token %
1	892 (30.56)	1270 (33.84)	33977 (65.38)	65.38
2	784 (26.86)	1023 (27.26)	10888 (20.95)	86.33
3	675 (23.12)	791 (21.08)	5457 (10.50)	96.83
4	282 (9.66)	293 (7.81)	802 (1.54)	98.37
5	134 (4.59)	134 (3.57)	265 (0.51)	98.88
6	67 (2.30)	70 (1.87)	99 (0.19)	99.07
7	31 (1.06)	31 (0.83)	42 (0.08)	99.15
8	23 (0.79)	23 (0.61)	29 (0.06)	99.21
9	12 (0.41)	12 (0.32)	22 (0.04)	99.25
10	5 (0.17)	5 (0.13)	5 (0.01)	99.26
11	6 (0.21)	6 (0.16)	10 (0.02)	99.28
12	2 (0.07)	2 (0.05)	2 (0.00)	99.29
13	1 (0.03)	1 (0.03)	1 (0.00)	99.29
14	3 (0.10)	3 (0.08)	3 (0.01)	99.29
15				
16				
17	2 (0.07)	2 (0.05)	2 (0.00)	99.29
18				
19				
20				
21				
22				
23				
24				
25				

Off-List:	??	87 (2.32)	365 (0.70)	99.99
Total	2919+?	3753 (100)	51969 (100)	100.00

Families List [↑]

Family [number of tokens]

BNC-COCA-1,000 Families: [fams 826 : types 1047 : tokens 16491]

able_[17] about_[34] above_[2] absolute_[4] accept_[2] across_[2] act_[15] actual_[3] add_[29] address_[18] admit_[2] advertise_[2] afford_[3] afraid_[5] afternoon_[4] again_[19] age_[29] ago_[34] agree_[13] ahead_[13] air_[28] all_[21] allow_[4] almost_[21] alone_[11] along_[11] already_[4] also_[32] always_[7] amount_[18] and_[6] angry_[6] animal_[5] answer_[17] any_[14] apart_[4] apparent_[3] appear_[3] area_[14] arm_[39] around_[33] arrange_[2] arrive_[6] art_[55] as_[26] ask_[13] aware_[6] away_[60] awful_[2] baby_[7] back_[101] bad_[36] bag_[9] ball_[7] bank_[16] base_[15] basic_[4] be_[167] bear_[2] beat_[6] beauty_[6] become_[40] bed_[15] before_[16] begin_[17] behind_[2] believe_[12] below_[11] best_[29] bet_[1] better_[43] between_[1] big_[31] bill_[12] birth_[8] bit_[12] black_[34] blood_[16] blow_[9] blue_[16] board_[17] boat_[2] body_[19] bone_[1] book_[21] born_[8] both_[1] bother_[3] bottle_[4] bottom_[4] box_[4] boy_[9] break_[36] breakfast_[2] breath_[10] bright_[8] bring_[25] brother_[14] brown_[8] build_[20] burn_[7] bus_[7] business_[20] busy_[5] but_[1] buy_[26] by_[10] call_[24] camp_[6] can_[88] car_[34] card_[14] care_[54] carry_[11] case_[15] cat_[2] catch_[15] cause_[18] centre_[13] certain_[20] chair_[9] chance_[14] change_[51] charge_[7] check_[5] chicken_[4] child_[53] choice_[5] choose_[3] church_[3] city_[18] class_[21] clean_[11] clear_[15] climb_[1] clock_[2] close_[56] closed_[2] clothes_[5] club_[2] coffee_[8] cold_[11] collect_[10] college_[42] colour_[8] come_[69] comfort_[12] company_[48] complete_[7] computer_[20] concern_[15] consider_[15] continue_[3] control_[19] conversation_[1] cook_[11] corner_[8] cost_[29] could_[60] count_[4] country_[23] couple_[21] course_[19] court_[23] cover_[18] crazy_[5] crime_[13] cross_[8] cry_[4] cup_[44] cut_[23] dad_[2] dance_[2] danger_[8] dark_[12] date_[2] daughter_[10] day_[72] dead_[8] deal_[19] death_[16] decide_[6] deep_[10] degree_[14] depend_[4] die_[23] difference_[16] different_[46] difficult_[20] dig_[2] dinner_[11] dirty_[2] discover_[4] do_[34] doctor_[14] dog_[4] door_[54] double_[2] doubt_[1] down_[111] draw_[8] dream_[5] dress_[8] drink_[16] drive_[34] drop_[7] drug_[33] dry_[5] ear_[2] early_[41] east_[13] easy_[23] eat_[16] edge_[2] educate_[64] egg_[10] either_[8] else_[10] employ_[8] end_[30] enjoy_[4] enough_[30] enter_[3] especially_[5] even_[12] evening_[2] ever_[31] every_[3] exact_[13] excite_[8] expect_[10] expensive_[10] experience_[15] explain_[10] express_[6] eye_[61] face_[31] fact_[3] fair_[2] fall_[32] family_[33] far_[41] farm_[4] fast_[21] fat_[18] father_[14] fear_[3] feel_[46] fellow_[2] field_[15] fight_[10] figure_[18] fill_[14] film_[3] final_[10] find_[59] fine_[6] finger_[10] finish_[9] fire_[16] first_[2] fish_[6] fit_[5] fix_[6] floor_[13] fly_[4] follow_[16] food_[24] foot_[48] football_[14] for_[5] force_[32] forest_[6] forget_[7] form_[13] forward_[18] free_[21] fresh_[20] friend_[19] from_[3] front_[16] full_[12] fun_[16] game_[32] gas_[9] general_[16] get_[126] girl_[10] give_[65] glad_[3] glance_[5] glass_[8] go_[48] gold_[10] good_[55] govern_[31] grandfather_[2] great_[15] green_[12] grey_[2] ground_[17] group_[53] grow_[28] guess_[7] gun_[8] guy_[12] hair_[43]

half_[8] hall_[2] hand_[77] handle_[7] hang_[11] happen_[23] happy_[13] hard_[49] hardly_[2] hat_[2] have_[57] head_[44] health_[64] hear_[30] heart_[18] heat_[31] heavy_[6] hell_[2] help_[48] here_[37] hide_[4] high_[81] history_[22] hit_[17] hold_[31] hole_[2] holiday_[2] home_[58] hope_[10] horse_[2] hospital_[10] hot_[16] hour_[50] house_[12] how_[102] however_[2] huge_[2] human_[32] hurry_[1] hurt_[8] husband_[5] ice_[4] idea_[12] imagine_[15] important_[48] in_[29] indeed_[3] inform_[37] inside_[4] insure_[12] interest_[56] internet_[2] involve_[9] issue_[48] job_[24] join_[14] joke_[4] judge_[6] jump_[5] just_[58] keep_[37] key_[12] kick_[6] kid_[19] kill_[12] kind_[16] kiss_[2] kitchen_[8] knock_[8] know_[34] lady_[6] land_[7] large_[41] last_[12] late_[52] laugh_[2] law_[53] lay_[14] lead_[37] learn_[30] least_[12] leave_[16] left_[17] leg_[9] less_[18] let_[33] letter_[8] level_[34] lie_[8] life_[56] lift_[9] light_[30] like_[16] line_[25] lip_[2] list_[9] listen_[8] little_[48] live_[30] load_[1] local_[34] lock_[4] long_[75] look_[65] lose_[22] lot_[29] loud_[2] love_[20] low_[32] luck_[6] lunch_[4] machine_[5] mad_[2] main_[6] major_[24] make_[194] man_[41] manage_[20] mark_[1] market_[22] marry_[26] master_[2] matter_[18] may_[49] maybe_[11] meal_[2] mean_[16] meet_[32] member_[32] mention_[7] mess_[1] middle_[16] might_[19] mile_[14] milk_[3] mind_[19] minute_[63] miss_[6] mistake_[7] moment_[18] money_[46] month_[45] more_[169] morning_[17] most_[85] mother_[24] mountain_[2] mouth_[14] move_[31] movie_[11] much_[36] mum_[2] music_[27] must_[5] name_[24] nation_[88] nature_[18] near_[12] necessary_[8] neck_[4] need_[22] neighbour_[2] never_[32] new_[118] news_[25] nice_[15] night_[35] no_[6] noise_[4] normal_[7] north_[9] nose_[2] note_[14] notice_[8] now_[21] number_[25] nurse_[4] obvious_[5] off_[67] offer_[15] office_[25] officer_[14] often_[13] oil_[31] ok_[6] old_[35] on_[20] once_[18] one_[10] only_[32] open_[36] orange_[2] order_[11] other_[91] ought_[2] out_[150] over_[36] own_[2] owned_[5] pack_[1] page_[10] pain_[5] paint_[2] pair_[2] paper_[16] parent_[10] park_[8] part_[23] particular_[8] party_[20] pass_[23] past_[12] pay_[30] people_[76] perfect_[1] perhaps_[10] person_[17] photograph_[3] pick_[4] picture_[12] piece_[12] place_[17] plan_[20] plant_[21] play_[78] please_[9] point_[34] police_[21] poor_[9] pop_[3] position_[4] possible_[20] post_[3] pound_[8] power_[23] prepare_[8] present_[11] press_[10] pretty_[14] price_[29] prison_[6] probably_[10] problem_[52] programme_[55] promise_[7] protect_[14] public_[50] pull_[25] push_[19] put_[30] quarter_[2] question_[24] quick_[24] quiet_[8] quite_[15] race_[12] radio_[16] rain_[4] raise_[32] rate_[47] rather_[3] reach_[16] read_[29] ready_[6] real_[21] realise_[8] really_[25] reason_[13] recent_[33] record_[15] red_[26] relate_[31] remember_[11] rent_[4] report_[32] responsible_[6] rest_[12] return_[12] rich_[6] rid_[2] ride_[7] right_[89] ring_[4] rise_[12] road_[15] rock_[4] roll_[15] room_[50] round_[1] rule_[15] run_[34] sad_[3] safe_[20] same_[1] save_[11] say_[45] school_[84] science_[25] sea_[3] seat_[18] second_[18] secure_[19] see_[43] seem_[9] sell_[28] send_[18] sense_[12] serious_[13] serve_[13] service_[64] set_[42] settle_[4] sex_[26] shake_[8] shall_[9] shape_[8] share_[20] shirt_[5] shoe_[6] shoot_[16] shop_[8] short_[19] should_[10] shoulder_[10] show_[51] shut_[13] sick_[6] side_[34] sight_[2] sign_[13] simple_[11] since_[3] sing_[2] single_[9] sir_[2] sister_[10] sit_[43] situation_[7] size_[10] skin_[4] sky_[6] sleep_[10] slight_[6] slip_[2] slow_[16] small_[26] smell_[1] smile_[10] smoke_[2] snow_[2] so_[53] soft_[6] some_[9] son_[11] song_[8] soon_[23] sorry_[4] sort_[6] sound_[14] south_[13] space_[13] speak_[9] special_[18] spend_[33] sport_[8] spot_[2] spring_[9] square_[3] stage_[8] stand_[32] star_[8] stare_[12] start_[25] state_[59] station_[16] stay_[22] step_[19] stick_[8] still_[11] stone_[4] stop_[8] store_[16] story_[29] straight_[13] strange_[4] street_[11] strike_[4] strong_[17] student_[110] study_[70] stuff_[9] subject_[6] sudden_[9] suggest_[23] suit_[8] summer_[12] sun_[4] support_[38] suppose_[3] sure_[22] surprise_[6] sweet_[2] swim_[2] system_[55] table_[16] take_[140] talk_[21] tall_[6] tape_[2] taste_[4] tax_[49] tea_[2] teach_[59] team_[36]

tear_[9] telephone_[28] television_[23] tell_[20] tend_[3] term_[17] terrible_[2] test_[20] thank_[21] that_[5] the_[3] then_[48] there_[27] thick_[2] thing_[65] think_[51] this_[2] throat_[2] through_[10] throw_[19] tie_[3] time_[64] tire_[5] today_[10] together_[23] tomorrow_[11] tonight_[7] too_[43] top_[16] total_[14] touch_[2] town_[10] track_[8] train_[18] treat_[23] tree_[14] trip_[6] trouble_[3] trust_[5] truth_[4] try_[22] turn_[42] two_[10] type_[6] under_[6] understand_[24] up_[128] use_[115] usual_[2] very_[109] video_[12] view_[11] visit_[4] voice_[17] wait_[20] wake_[4] walk_[31] wall_[10] want_[30] war_[23] warm_[6] wash_[5] waste_[7] watch_[23] water_[59] wave_[4] way_[55] wear_[32] weather_[4] web_[2] week_[31] weight_[12] well_[19] west_[7] wheel_[1] when_[30] where_[28] while_[8] white_[28] whole_[23] why_[25] wide_[18] wife_[11] wild_[2] will_[175] win_[32] wind_[6] window_[23] wine_[10] wish_[5] woman_[28] wonder_[18] word_[20] work_[44] world_[51] worry_[5] would_[71] write_[32] wrong_[12] yard_[14] year_[121] yellow_[2] yesterday_[1] yet_[1] young_[38]

BNC-COCA-2,000 Families: [fams 526 : types 612 : tokens 3611]

access_[11] accident_[2] account_[5] active_[29] adult_[8] advance_[4] advantage_[2] advice_[5] advise_[5] affair_[8] affect_[15] agent_[8] aid_[4] alcohol_[6] alive_[6] announce_[6] apartment_[4] appeal_[2] apply_[2] appoint_[2] approach_[10] argue_[7] arrest_[2] article_[13] aside_[6] asleep_[2] assist_[8] associate_[4] assume_[7] attack_[4] attempt_[2] attend_[13] attention_[25] attitude_[9] attract_[6] available_[17] average_[15] avoid_[6] bake_[6] balance_[8] band_[2] bare_[4] basis_[6] battle_[4] bean_[2] beer_[2] bell_[4] belong_[1] belt_[2] bend_[6] benefit_[16] bike_[4] bite_[4] blame_[2] block_[4] blonde_[2] boil_[6] bomb_[4] bond_[4] borrow_[2] bounce_[2] bow_[2] bowl_[6] branch_[4] brand_[4] breast_[6] breathe_[2] brick_[2] broad_[2] brush_[2] buck_[2] butter_[8] button_[4] cable_[6] calm_[2] camera_[6] cap_[2] capital_[12] career_[4] cash_[2] cast_[4] cent_[1] century_[19] chain_[2] challenge_[9] champion_[10] character_[2] cheek_[4] cheese_[4] chest_[2] chief_[15] chop_[11] cigarette_[4] circle_[4] citizen_[7] claim_[2] classic_[2] clip_[2] coach_[14] combine_[2] comment_[4] commerce_[4] commit_[10] committee_[2] common_[21] community_[34] compare_[5] complicate_[2] concentrate_[2] condition_[6] connect_[2] contact_[6] contain_[8] contract_[2] contribute_[13] convince_[2] copy_[2] correct_[2] council_[4] county_[2] crack_[2] cream_[7] create_[14] credit_[8] criminal_[8] culture_[27] curl_[2] current_[11] customer_[2] damage_[3] debt_[2] decision_[13] defence_[16] deliver_[3] demand_[3] department_[8] depress_[2] describe_[10] desert_[2] design_[8] desk_[2] desperate_[2] destroy_[2] detail_[3] detect_[1] determine_[9] develop_[55] dine_[4] direct_[2] directed_[19] direction_[12] disappear_[3] discuss_[13] disease_[14] distance_[6] district_[6] dollar_[6] drama_[6] due_[2] earn_[6] economy_[50] edit_[6] effect_[17] effort_[4] elder_[4] elect_[19] emotion_[2] encourage_[6] energy_[10] engage_[4] enormous_[2] entire_[14] environment_[27] equal_[6] equipment_[2] establish_[1] estate_[4] event_[2] eventually_[5] evidence_[18] evil_[2] examine_[8] example_[12] exchange_[4] exist_[4] extend_[4] extreme_[4] fail_[1] fairy_[1] faith_[3] familiar_[2] famous_[3] fan_[2] feature_[3] female_[12] file_[8] finance_[30] firm_[9] flag_[3] flash_[2] flight_[2] flow_[4] fold_[2] folk_[2] foreign_[21] forth_[11] fortune_[4] frame_[2] frankly_[2] fruit_[4] fry_[2] fund_[27] future_[7] gain_[6] gate_[4] gather_[6] gay_[7] generation_[6] gentleman_[2] gift_[10] goal_[11] golf_[4] grab_[4] grade_[8] grand_[2] grant_[2] grocer_[2] guard_[6] guest_[2] guilty_[6] hedge_[1] hire_[1] hook_[1] hotel_[4] identify_[13] ignore_[1] ill_[4] illustrate_[3] image_[4] immediate_[5] improve_[10] inch_[3] include_[16] income_[13] increase_[28] indicate_[15] individual_[10] industry_[7] influence_[8] injure_[2] innocent_[2] instruct_[2] instrument_[4] intense_[4] interview_[5] introduce_[2] investigate_[4]

item_[4] jacket_[2] jeans_[2] juice_[11] junior_[8] justice_[6] knee_[6] knowledge_[4] labour_[8] language_[12] lawyer_[4] league_[10] lean_[8] legal_[12] length_[2] lesson_[5] library_[2] licence_[4] likely_[22] limit_[3] loan_[6] loss_[8] magazine_[12] mail_[4] male_[10] map_[4] mass_[6] mate_[2] material_[8] mathematics_[2] measure_[6] meat_[2] medical_[18] medicine_[2] melt_[2] mental_[8] message_[6] metal_[2] military_[27] minister_[4] minor_[12] mirror_[2] mix_[1] model_[9] modern_[4] moon_[2] murder_[4] narrow_[1] native_[4] nervous_[6] newspaper_[6] nowhere_[1] oak_[2] object_[1] observe_[2] occasion_[2] occur_[12] official_[33] onion_[12] operate_[8] opinion_[4] opportunity_[18] oppose_[2] opposite_[6] option_[4] ordinary_[2] organize_[24] otherwise_[5] oven_[2] pan_[2] partner_[2] path_[3] patient_[14] pattern_[2] pause_[2] peace_[11] pension_[2] percent_[36] perform_[10] period_[12] physical_[18] piano_[2] pile_[1] pine_[2] plain_[2] plane_[5] plastic_[3] plate_[2] pocket_[2] poem_[2] policy_[38] politics_[51] pollute_[2] pool_[2] popular_[13] population_[11] positive_[12] potato_[3] pour_[4] practise_[6] prefer_[3] pregnant_[4] president_[21] pressure_[10] prevent_[5] previous_[8] pride_[2] prime_[4] private_[17] process_[17] produce_[4] product_[29] profession_[5] progress_[2] project_[5] property_[12] propose_[2] proud_[3] prove_[7] provide_[54] pump_[1] punish_[2] purpose_[8] quality_[14] quit_[2] quote_[1] range_[8] rapid_[4] react_[2] receive_[21] recognize_[2] recommend_[2] reduce_[19] refer_[3] region_[3] regular_[4] release_[4] rely_[2] remain_[6] remark_[2] remove_[4] repeat_[2] represent_[5] require_[7] research_[56] reserve_[1] resist_[1] restaurant_[2] result_[32] retire_[2] rip_[2] risk_[31] role_[24] row_[2] royal_[2] rush_[2] sale_[5] salt_[24] sauce_[3] scale_[4] scene_[2] score_[18] screen_[7] screw_[1] search_[8] season_[16] secret_[4] section_[2] seek_[5] select_[5] senior_[20] sentence_[2] series_[5] shadow_[2] shed_[2] sheet_[5] shine_[4] shore_[1] shower_[2] signal_[2] silence_[2] silver_[2] similar_[14] site_[5] ski_[1] skill_[21] slide_[2] smart_[4] social_[74] society_[10] soldier_[3] solid_[2] southern_[2] species_[5] specific_[5] speech_[7] speed_[3] spell_[2] spin_[2] spirit_[2] split_[1] spread_[4] stable_[2] staff_[8] standard_[15] steel_[1] stir_[9] stock_[18] storm_[2] strength_[2] stretch_[2] style_[2] success_[15] suffer_[5] sugar_[14] supply_[4] surface_[2] survive_[6] sweep_[3] swing_[2] switch_[3] tale_[3] tank_[2] technology_[21] teenage_[2] theatre_[2] therefore_[3] thin_[1] threat_[4] thus_[2] ticket_[5] title_[2] tomato_[2] tone_[4] tool_[2] topic_[12] tough_[8] tour_[1] towel_[2] trace_[2] trade_[13] tradition_[4] traffic_[2] trial_[4] truck_[8] trunk_[2] tune_[3] union_[2] unite_[1] university_[4] upper_[2] upset_[4] value_[14] various_[5] vary_[8] vegetable_[4] version_[2] victim_[2] violent_[2] vote_[14] wage_[3] weak_[2] weapon_[12] welcome_[6] western_[6] wing_[2] wipe_[2] wise_[2] wrap_[6]

BNC-COCA-3,000 Families: [fams 340 : types 365 : tokens 1527]

abort_[4] abuse_[10] academy_[6] accomplish_[1] accountable_[1] accurate_[4] achieve_[10] acre_[1] adjust_[2] administration_[14] administrator_[4] adopt_[2] affirm_[1] agency_[23] agenda_[2] aggressive_[2] airline_[1] album_[2] alien_[2] amend_[2] analyse_[18] analyst_[4] annual_[6] appropriate_[4] approve_[2] approximate_[1] arise_[2] armed_[4] aspect_[8] assault_[2] assembly_[2] assess_[2] assumption_[2] athlete_[4] author_[3] authority_[7] award_[2] behave_[4] behaviour_[18] belief_[4] bench_[2] biological_[2] border_[4] budget_[16] bureau_[2] campaign_[17] cancer_[16] candidate_[8] carbon_[1] carve_[2] category_[2] catholic_[2] celebrate_[2] cell_[11] chairman_[6] charter_[2] chemical_[2] civil_[11] civilian_[2] climate_[2] clinic_[2] code_[2] colleague_[2] column_[2] compete_[3] complaint_[2] complex_[2] concept_[10] conclusion_[6] conduct_[14] confer_[12] confident_[2] conflict_[4] congress_[3] conservative_[2] consistent_[2] constitution_[4] construct_[3]

consult_[1] consume_[4] contemporary_[2] context_[2] convention_[2] cooperate_[2] counsel_[8] craft_[2] crew_[2] crisis_[4] criteria_[4] critic_[3] curriculum_[2] cycle_[2] data_[26] debate_[2] decade_[10] decline_[2] deficit_[6] define_[1] democrat_[12] demonstrate_[4] deputy_[2] destruction_[4] device_[2] devote_[2] digital_[2] disabled_[8] disagree_[2] disaster_[2] distinct_[2] diverse_[2] document_[15] domestic_[8] eastern_[4] effective_[12] efficient_[5] element_[3] eliminate_[2] emerge_[2] emergency_[4] emit_[2] emphasis_[2] enable_[1] enforce_[8] ensure_[1] era_[2] error_[2] essay_[2] essential_[1] ethnic_[8] exclusive_[2] executive_[17] expert_[2] external_[2] facility_[2] factor_[21] faculty_[6] fade_[2] failure_[4] federal_[41] fee_[2] fibre_[6] fiction_[2] flexible_[2] focus_[12] foster_[2] frequent_[6] fuel_[5] funeral_[2] gallery_[2] gang_[2] gap_[2] gender_[8] geography_[1] gesture_[2] global_[7] graduate_[18] gross_[2] guitar_[2] hazard_[1] heel_[2] hip_[2] holy_[2] host_[6] household_[2] humour_[2] hypothesis_[2] immigrant_[2] immune_[2] impact_[7] implement_[2] impression_[4] independent_[6] index_[2] infect_[1] inflate_[2] insight_[2] institution_[7] intellectual_[2] intelligence_[4] interact_[2] internal_[4] international_[27] invest_[15] jail_[1] joint_[4] jury_[2] leather_[2] legislate_[4] liberal_[2] liberty_[2] likeness_[2] literature_[2] magnet_[2] majority_[8] margin_[2] media_[12] medium_[3] method_[5] minimum_[2] missile_[2] mixture_[2] modify_[6] mortal_[2] motion_[6] multiple_[1] museum_[2] mutual_[2] negative_[6] net_[4] network_[8] nod_[5] novel_[2] nuclear_[8] objective_[2] obtain_[3] offence_[2] offend_[2] opera_[2] oral_[2] oriented_[2] overwhelm_[2] pace_[2] palm_[4] participant_[10] participate_[4] passenger_[2] pave_[1] peer_[1] penalty_[2] pepper_[15] permission_[2] permit_[2] personnel_[2] perspective_[2] phrase_[2] poll_[7] pose_[6] potential_[2] poverty_[2] powder_[4] precede_[1] predict_[5] primary_[4] principal_[2] priority_[4] prize_[2] procedure_[2] professor_[14] profit_[8] prominent_[2] prosecute_[4] protein_[6] psychology_[6] publish_[10] racial_[2] raw_[2] receiver_[2] reflect_[4] reform_[14] refuge_[2] regulate_[6] relative_[10] religious_[14] remote_[2] republic_[4] reside_[2] resolution_[2] resolve_[4] resource_[18] respond_[6] response_[4] reveal_[1] revenue_[5] review_[4] route_[2] rural_[6] sacrifice_[2] sample_[10] sanction_[1] satisfaction_[2] secretary_[6] sector_[4] senate_[5] sensitive_[2] severe_[2] significant_[32] silent_[4] slice_[1] software_[6] solution_[4] solve_[4] sophisticated_[2] source_[15] squeeze_[2] stain_[1] stake_[2] statistic_[3] status_[1] stem_[2] strategy_[5] structure_[2] studio_[2] substance_[4] subtle_[2] suicide_[4] sum_[1] supreme_[1] surgery_[3] survey_[10] sustain_[2] task_[4] technique_[4] temperature_[6] tennis_[4] terror_[8] text_[5] theme_[2] theory_[2] trail_[2] transition_[2] transport_[2] tremendous_[2] trigger_[1] troop_[4] undergo_[2] uniform_[2] urban_[4] variety_[4] vast_[2] venture_[4] vessel_[2] vice_[3] violate_[6] violence_[2] virtual_[2] vision_[2] visual_[3] vulnerable_[4] weigh_[3] welfare_[8]

BNC-COCA-4,000 Families: [fams 74 : types 74 : tokens 193]

acid_[2] adverse_[1] anniversary_[2] attorney_[11] automobile_[2] baseball_[10] boarder_[1] bulb_[2] calorie_[4] campus_[2] cholesterol_[5] chronic_[2] comic_[1] compel_[2] consistency_[2] convenience_[2] copyright_[1] cord_[1] couch_[2] crude_[1] debut_[2] dim_[1] dioxide_[1] elementary_[6] enrol_[2] eyebrow_[2] fiscal_[2] flour_[4] fossil_[1] garlic_[8] glimpse_[2] greenhouse_[1] harass_[2] hardware_[2] hostage_[2] hug_[1] immigrate_[2] impair_[1] indigenous_[2] integral_[1] legislature_[2] lemon_[6] lesbian_[1] lung_[2] medal_[4] metropolitan_[2] milligram_[2] monetary_[2] monument_[2] olive_[3] patrol_[2] pill_[2] plead_[2] prescription_[2] prop_[1] regress_[1] sip_[1] slam_[2] sleeve_[1] soak_[1] soap_[2] sodium_[5] solar_[2] spine_[1] spit_[2] stadium_[2] steer_[1] tablespoon_[23] tag_[2] telescope_[2] tilt_[1] tobacco_[4] transcript_[2] wagon_[2]

BNC-COCA-5,000 Families: [fams 35 : types 35 : tokens 64]

advocacy_[1] aide_[1] aisle_[2] aloud_[1] bail_[1] basketball_[12] bulletin_[1] carbohydrate_[7] cellular_[1] clap_[1] clasp_[1] cock_[1] comb_[1] deviate_[1] fend_[1] gram_[3] intercourse_[1] lime_[1] mall_[2] medication_[2] oval_[1] pharmaceutical_[1] pickup_[2] porch_[2] precaution_[1] saturate_[1] scoop_[1] serial_[1] shuttle_[2] simmer_[2] sour_[1] toll_[2] undergraduate_[2] vacation_[2] vinegar_[1]

BNC-COCA-6,000 Families: [fams 17 : types 17 : tokens 18]

bachelor_[1] clove_[1] colon_[1] conjure_[1] focal_[1] freak_[1] freelance_[1] genome_[1] locker_[1] mash_[1] mince_[1] pant_[1] payroll_[1] peanut_[2] rebound_[1] transcribe_[1] vain_[1]

BNC-COCA-7,000 Families: [fams 9 : types 9 : tokens 9]

amino_[1] broth_[1] cinnamon_[1] ethic_[1] hispanic_[1] marrow_[1] prostate_[1] vanilla_[1] vantage_[1]

BNC-COCA-8,000 Families: [fams 2 : types 2 : tokens 3]

freshman_[2] soy_[1]

BNC-COCA-9,000 Families: [fams 1 : types 1 : tokens 2]

playoff_[2]

BNC-COCA-10,000 Families: [fams 1 : types 1 : tokens 1]

supra_[1]

BNC-COCA-11,000 Families: [fams 3 : types 3 : tokens 4]

boomer_[1] quo_[1] skillet_[2]

BNC-COCA-12,000 Families: [fams : types : tokens]

BNC-COCA-13,000 Families: [fams : types : tokens]

BNC-COCA-14,000 Families: [fams : types : tokens]

BNC-COCA-15,000 Families: [fams : types : tokens]

BNC-COCA-16,000 Families: [fams : types : tokens]

BNC-COCA-17,000 Families: [fams : types : tokens]

BNC-COCA-18,000 Families: [fams : types : tokens]

BNC-COCA-19,000 Families: [fams : types : tokens]

BNC-COCA-20,000 Families: [fams : types : tokens]

BNC-COCA-21,000 Families: [fams : types : tokens]

BNC-COCA-22,000 Families: [fams : types : tokens]

BNC-COCA-23,000 Families: [fams : types : tokens]

BNC-COCA-24,000 Families: [fams : types : tokens]

BNC-COCA-25,000 Families: [fams : types : tokens]

OFFLIST: [?: types 40 : tokens 128]

african_[2] airport_[2] american_[30] arab_[3] bathroom_[1] bedroom_[2] birthday_[4] bitch_[1] british_[1] chinese_[1] classroom_[6] doorway_[2] english_[2] european_[2] forever_[3] french_[1] hallway_[2] headline_[2] homeland_[2] iraqi_[3] israeli_[1] jewish_[1] lawsuit_[2] longtime_[2] muslim_[1] nongovernmental_[1] olympic_[2] online_[2] palestinian_[1] pm_[1] preservice_[1]

proofread_[1] someday_[1] spokesman_[2] spokeswoman_[1] teaspoon_[30] touchdown_[2] upstairs_[1] weekend_[2]

Appendix 4. Cut-off of one occurrence per million tokens resulting in 3,006 word families plus off-list types

Word List	Families (%)	Types (%)	Tokens (%)	Cumul. token %
1	892 (30.56)	1270 (33.84)	33977 (65.38)	65.38
2	784 (26.86)	1023 (27.26)	10888 (20.95)	86.33
3	675 (23.12)	791 (21.08)	5457 (10.50)	96.83
4	282 (9.66)	293 (7.81)	802 (1.54)	98.37
5	134 (4.59)	134 (3.57)	265 (0.51)	98.88
6	67 (2.30)	70 (1.87)	99 (0.19)	99.07
7	31 (1.06)	31 (0.83)	42 (0.08)	99.15
8	23 (0.79)	23 (0.61)	29 (0.06)	99.21
9	12 (0.41)	12 (0.32)	22 (0.04)	99.25
10	5 (0.17)	5 (0.13)	5 (0.01)	99.26
11	6 (0.21)	6 (0.16)	10 (0.02)	99.28
12	2 (0.07)	2 (0.05)	2 (0.00)	99.29
13	1 (0.03)	1 (0.03)	1 (0.00)	99.29
14	3 (0.10)	3 (0.08)	3 (0.01)	99.29
15				
16				
17	2 (0.07)	2 (0.05)	2 (0.00)	99.29
18				
19				
20				
21				
22				
23				
24				

25

Off-List:	??	87 (2.32)	365 (0.70)	99.99
Total	2919+?	3753 (100)	51969 (100)	≈100.00

Families List [↑]

Family [number of tokens]

BNC-COCA-1,000 Families: [fams 892 : types 1270 : tokens 33977]

able_[40] about_[56] above_[9] absolute_[11] accept_[25] across_[6] act_[41] actual_[9] add_[87] address_[33] admit_[4] advertise_[7] afford_[9] afraid_[7] after_[1] afternoon_[16] again_[37] age_[47] ago_[40] agree_[29] ahead_[26] air_[64] all_[23] allow_[12] almost_[39] alone_[14] along_[17] already_[9] also_[60] always_[9] amaze_[7] amount_[48] and_[7] angry_[9] animal_[18] answer_[35] any_[27] apart_[19] apparent_[4] appear_[12] area_[56] arm_[77] around_[56] arrange_[3] arrive_[25] art_[104] as_[29] ashamed_[1] ask_[21] at_[1] aunt_[2] aware_[18] away_[84] awful_[3] baby_[23] back_[176] bad_[67] bag_[30] ball_[19] bank_[33] bar_[8] base_[55] basic_[26] bath_[2] be_[176] beach_[3] bear_[25] beat_[19] beauty_[18] become_[73] bed_[31] before_[26] begin_[44] behind_[12] believe_[24] below_[16] best_[78] bet_[5] better_[68] between_[1] big_[56] bill_[25] bird_[4] birth_[21] bit_[32] black_[89] blood_[37] blow_[17] blue_[45] board_[43] boat_[6] body_[42] bone_[7] book_[44] boring_[1] born_[13] both_[3] bother_[8] bottle_[14] bottom_[5] box_[15] boy_[17] bread_[9] break_[84] breakfast_[10] breath_[16] bright_[29] bring_[40] brother_[18] brown_[25] build_[69] burn_[12] bus_[15] business_[53] busy_[10] but_[1] buy_[54] by_[14] cake_[5] call_[42] camp_[18] can_[154] car_[77] card_[29] care_[99] carry_[27] case_[53] cat_[2] catch_[24] cause_[45] centre_[62] certain_[40] chair_[29] chance_[26] change_[117] charge_[30] cheap_[2] check_[22] chicken_[12] child_[98] chip_[10] choice_[17] choose_[11] church_[15] city_[36] class_[51] clean_[15] clear_[44] climb_[20] clock_[6] close_[117] closed_[3] clothes_[18] club_[10] coat_[9] coffee_[16] cold_[19] collect_[32] college_[69] colour_[28] come_[104] comfort_[17] company_[102] complete_[32] computer_[43] concern_[43] consider_[31] continue_[20] control_[54] conversation_[7] cook_[29] cool_[20] corner_[20] cost_[75] could_[85] count_[7] country_[39] couple_[25] course_[30] court_[44] cover_[52] crazy_[9] crime_[29] cross_[24] cry_[19] cup_[103] cut_[44] dad_[2] dance_[8] danger_[22] dark_[47] date_[8] daughter_[17] day_[132] dead_[20] deal_[39] dear_[2] death_[43] decide_[11] deep_[29] definite_[4] degree_[24] depend_[6] die_[46] difference_[47] different_[93] difficult_[52] dig_[8] dinner_[20] dirty_[2] discover_[9] do_[43] doctor_[23] dog_[13] door_[93] double_[12] doubt_[14] down_[182] draw_[27] dream_[18] dress_[26] drink_[35] drive_[51] drop_[25] drug_[75] dry_[17] ear_[14] early_[78] earth_[3] east_[22] easy_[45] eat_[38] edge_[16] educate_[156] egg_[23] either_[8] else_[15] employ_[33] empty_[10] end_[67] engine_[6] enjoy_[4] enough_[45] enter_[11] especially_[9] even_[17] evening_[16] ever_[39] every_[11] exact_[25] excite_[13] excuse_[2] expect_[16] expensive_[10] experience_[59] explain_[24] express_[21] extra_[15] eye_[112] face_[105]

fact_[15] fair_[13] fall_[63] family_[61] far_[86] farm_[14] fast_[37] fat_[32] father_[23] favourite_[5] fear_[12] feel_[94] fellow_[8] field_[28] fight_[23] figure_[25] fill_[25] film_[23] final_[39] find_[109] fine_[11] finger_[33] finish_[21] fire_[37] first_[6] fish_[22] fit_[15] fix_[12] flat_[4] floor_[43] flower_[8] fly_[22] follow_[50] food_[77] foot_[107] football_[26] for_[5] force_[55] forest_[18] forget_[14] form_[31] fortunate_[2] forward_[20] four_[1] free_[73] fresh_[38] friend_[28] from_[3] front_[42] full_[35] fun_[21] game_[75] garden_[12] gas_[21] general_[33] gentle_[3] get_[186] girl_[21] give_[145] glad_[5] glance_[17] glass_[35] go_[60] gold_[24] good_[80] govern_[82] grandfather_[10] grass_[4] great_[56] green_[32] grey_[12] ground_[47] group_[108] grow_[81] guess_[7] gun_[28] guy_[20] hair_[62] half_[8] hall_[12] hand_[129] handle_[19] hang_[21] happen_[33] happy_[16] hard_[79] hardly_[8] hat_[10] hate_[5] have_[84] head_[108] health_[130] hear_[65] heart_[41] heat_[57] heavy_[19] hell_[9] help_[85] here_[54] hide_[13] high_[194] hill_[4] history_[67] hit_[38] hold_[60] hole_[17] holiday_[4] home_[109] honest_[1] honour_[3] hope_[23] horrible_[2] horse_[7] hospital_[27] hot_[43] hour_[78] house_[53] how_[168] however_[4] huge_[6] human_[77] hunt_[5] hurry_[1] hurt_[15] husband_[15] i_[6] ice_[13] idea_[24] imagine_[20] important_[69] in_[64] indeed_[3] inform_[101] inside_[14] instead_[6] insure_[38] interest_[96] internet_[6] involve_[31] island_[2] issue_[113] job_[48] join_[30] joke_[4] judge_[19] jump_[15] just_[96] keep_[70] key_[30] kick_[14] kid_[29] kill_[42] kind_[30] kiss_[10] kitchen_[20] knock_[11] know_[43] lady_[6] lake_[4] land_[38] large_[101] last_[26] late_[90] laugh_[14] law_[105] lay_[29] lead_[100] learn_[67] least_[19] leave_[29] left_[30] leg_[26] less_[33] let_[49] letter_[20] level_[89] lie_[18] life_[86] lift_[20] light_[81] like_[24] line_[54] lip_[17] list_[22] listen_[18] little_[79] live_[56] load_[6] local_[75] lock_[9] long_[124] look_[108] lose_[42] lot_[48] loud_[16] love_[31] low_[66] luck_[16] lunch_[17] machine_[9] mad_[6] main_[26] major_[49] make_[288] man_[77] manage_[82] mark_[14] market_[80] marry_[39] master_[6] matter_[36] may_[80] maybe_[16] meal_[11] mean_[43] meet_[54] member_[63] mention_[13] mess_[4] middle_[40] might_[32] mile_[38] milk_[13] mind_[43] minute_[109] miss_[9] mistake_[9] moment_[29] money_[75] month_[74] more_[248] morning_[41] most_[169] mother_[30] mountain_[13] mouth_[29] move_[46] movie_[15] much_[52] mum_[5] music_[71] must_[19] name_[41] nation_[170] nature_[50] near_[30] necessary_[17] neck_[18] need_[62] neighbour_[20] never_[56] new_[238] news_[40] nice_[18] night_[56] no_[9] noise_[6] normal_[11] north_[21] nose_[7] note_[34] notice_[12] now_[26] number_[56] nurse_[5] obvious_[18] odd_[3] off_[131] offer_[55] office_[55] officer_[30] often_[32] oil_[76] ok_[12] old_[47] on_[23] once_[20] one_[24] only_[59] open_[75] orange_[2] order_[29] other_[149] ought_[4] out_[232] over_[68] own_[25] owned_[11] pack_[10] page_[29] pain_[16] paint_[27] pair_[4] paper_[35] pardon_[1] parent_[26] park_[19] part_[54] particular_[20] party_[47] pass_[43] past_[35] pay_[95] people_[117] perfect_[9] perhaps_[15] person_[57] photograph_[8] pick_[8] picture_[26] piece_[39] place_[42] plan_[61] plant_[38] play_[119] please_[25] point_[59] police_[43] poor_[27] pop_[8] position_[18] possible_[51] post_[4] pot_[9] pound_[12] power_[63] prepare_[24] present_[35] press_[26] pretty_[33] price_[70] prison_[29] probably_[11] problem_[93] programme_[131] promise_[12] proper_[1] protect_[35] public_[120] pull_[46] push_[29] put_[50] quarter_[9] question_[57] quick_[53] quiet_[14] quite_[31] race_[34] radio_[33] rain_[16] raise_[42] rate_[92] rather_[3] reach_[38] read_[54] ready_[12] real_[50] realise_[18] really_[37] reason_[30] recent_[55] record_[29] red_[56] relate_[77] remember_[11] rent_[12] report_[82] responsible_[21] rest_[17] return_[25] rich_[10] rid_[4] ride_[22] right_[162] ring_[16] rise_[34] river_[7] road_[44] rock_[20] roll_[34] room_[85] rough_[3] round_[13] rule_[52] run_[61] sad_[9] safe_[43] same_[2] save_[18] say_[66] scare_[7] school_[133] science_[60] sea_[14] seat_[27] second_[30] secure_[48] see_[71] seem_[40] self_[4] sell_[48] send_[32] sense_[30]

serious_[43] serve_[60] service_[129] set_[75] settle_[10] sex_[72] shake_[8] shall_[13] shape_[15] share_[40] ship_[2] shirt_[20] shoe_[14] shoot_[32] shop_[26] short_[37] should_[14] shoulder_[34] shout_[1] show_[83] shut_[16] sick_[13] side_[67] sight_[11] sign_[33] simple_[31] since_[3] sing_[15] single_[20] sir_[3] sister_[12] sit_[83] situation_[34] size_[22] skin_[16] sky_[21] sleep_[23] slight_[12] slip_[12] slow_[35] small_[69] smell_[4] smile_[28] smoke_[13] snow_[6] so_[89] soft_[12] some_[26] son_[19] song_[17] soon_[37] sorry_[4] sort_[11] sound_[33] south_[20] space_[47] speak_[25] special_[58] spend_[73] sport_[37] spot_[7] spring_[21] square_[7] stage_[14] stand_[70] star_[26] stare_[25] start_[49] state_[121] station_[28] stay_[44] steal_[6] step_[46] stick_[14] still_[15] stone_[6] stop_[28] store_[43] story_[47] straight_[38] strange_[12] street_[23] strike_[20] strong_[51] student_[192] study_[161] stuff_[9] stupid_[5] subject_[16] sudden_[26] suggest_[31] suit_[33] summer_[31] sun_[20] support_[94] suppose_[7] sure_[34] surprise_[21] sweet_[3] swim_[2] system_[108] table_[33] take_[203] talk_[36] tall_[19] tape_[10] taste_[14] tax_[102] tea_[12] teach_[141] team_[60] tear_[30] telephone_[69] television_[43] tell_[25] tend_[6] term_[40] terrible_[8] test_[52] thank_[36] that_[10] the_[3] then_[98] there_[28] thick_[5] thing_[103] think_[83] this_[2] throat_[5] through_[16] throw_[37] tie_[18] tight_[8] time_[98] tire_[15] today_[23] together_[51] tomorrow_[18] tonight_[17] too_[62] tooth_[7] top_[47] total_[28] touch_[20] town_[34] track_[22] train_[45] travel_[16] treat_[60] tree_[41] trip_[15] trouble_[12] true_[2] trust_[9] truth_[10] try_[36] turn_[62] two_[14] type_[14] uncle_[2] under_[11] understand_[48] up_[171] use_[277] usual_[8] very_[195] video_[26] view_[24] visit_[23] voice_[38] wait_[27] wake_[10] walk_[68] wall_[42] want_[48] war_[42] warm_[20] wash_[14] waste_[18] watch_[38] water_[143] wave_[23] way_[91] wear_[67] weather_[10] web_[2] wed_[4] week_[59] weight_[27] well_[39] west_[19] wet_[1] wheel_[11] when_[47] where_[35] while_[17] white_[81] whole_[47] why_[33] wide_[37] wife_[19] wild_[7] will_[198] win_[74] wind_[24] window_[49] wine_[21] winter_[17] wish_[9] woman_[49] wonder_[33] wood_[4] word_[35] work_[111] world_[62] worry_[11] worse_[1] worth_[5] would_[116] write_[77] wrong_[23] yard_[26] year_[197] yellow_[16] yesterday_[7] yet_[9] you_[1] young_[52] zero_[2]

BNC-COCA-2,000 Families: [fams 784 : types 1023 : tokens 10888]

access_[37] accident_[8] account_[26] accuse_[2] active_[67] adapt_[2] adult_[11] advance_[11] advantage_[10] advice_[19] advise_[16] affair_[19] affect_[38] agent_[16] aid_[24] alarm_[6] alcohol_[14] alive_[13] alter_[1] amuse_[1] announce_[17] anxious_[1] apartment_[20] appeal_[4] apple_[6] apply_[17] appoint_[2] appreciate_[14] approach_[20] argue_[15] army_[4] arrest_[12] article_[39] aside_[13] asleep_[2] assist_[38] associate_[22] assume_[9] assure_[2] atmosphere_[2] attack_[16] attempt_[5] attend_[25] attention_[51] attitude_[17] attract_[17] automatic_[2] available_[42] average_[44] avoid_[12] awake_[4] background_[12] bake_[24] balance_[20] band_[12] bang_[2] bare_[9] bark_[1] basis_[8] bat_[3] battle_[8] bay_[4] bean_[6] beef_[2] beer_[8] beg_[3] bell_[6] belong_[7] belt_[2] bend_[8] benefit_[43] bike_[6] bind_[2] bite_[6] blame_[3] blind_[4] block_[11] blonde_[3] boil_[12] bomb_[15] bond_[7] boom_[2] boot_[7] borrow_[2] bounce_[2] bow_[2] bowl_[26] brain_[12] branch_[6] brand_[4] breast_[11] breathe_[9] breed_[1] brick_[6] bridge_[11] brief_[5] broad_[8] brush_[10] buck_[3] bucket_[2] bunch_[5] burst_[4] bury_[2] butter_[20] button_[6] cable_[12] cage_[2] calculate_[2] calm_[4] camera_[16] cap_[6] capable_[4] capital_[26] captain_[2] career_[30] carpet_[2] cart_[2] cash_[6] cast_[10] casual_[2] ceiling_[7] cent_[2] century_[30] chain_[8] challenge_[22] champion_[23] channel_[2] chapter_[2] character_[8]

chase_[2] chat_[1] cheek_[6] cheer_[1] cheese_[9] chest_[8] chief_[31] chocolate_[7] chop_[26] cigarette_[9] circle_[10] circumstance_[2] citizen_[19] claim_[8] classic_[4] clip_[4] cloud_[9] clue_[2] coach_[27] coal_[2] coast_[6] combine_[10] command_[15] comment_[9] commerce_[20] commit_[12] committee_[13] common_[44] community_[72] compare_[15] competition_[6] complain_[2] complicate_[6] concentrate_[4] condition_[39] connect_[4] conscious_[2] constant_[4] contact_[15] contain_[18] contract_[12] contribute_[22] convince_[7] cop_[2] cope_[7] copy_[7] correct_[7] council_[4] counter_[2] county_[5] cow_[2] crack_[4] crash_[8] cream_[15] create_[38] creature_[2] credit_[26] creep_[1] criminal_[24] crowd_[8] culture_[75] cure_[1] curious_[2] curl_[2] current_[38] customer_[9] damage_[11] dare_[4] debt_[8] decision_[27] decorate_[1] defence_[51] deliver_[13] demand_[9] deny_[6] department_[13] depress_[5] describe_[22] desert_[4] deserve_[7] design_[38] desire_[4] desk_[11] desperate_[6] destroy_[5] detail_[19] detect_[3] determine_[17] develop_[129] diet_[5] dine_[6] direct_[26] directed_[61] direction_[32] disappear_[4] disappoint_[2] discipline_[2] discuss_[37] disease_[31] dish_[8] distance_[14] district_[13] divide_[2] divorce_[4] dollar_[23] drag_[7] drama_[12] drawer_[4] due_[7] dust_[3] duty_[4] earn_[19] ease_[1] economy_[116] edit_[27] effect_[50] effort_[26] elder_[6] elect_[50] electric_[14] emotion_[15] encourage_[9] energy_[49] engage_[9] engineer_[16] enormous_[2] entertain_[2] entire_[31] environment_[84] equal_[17] equipment_[10] escape_[7] establish_[14] estate_[10] event_[20] eventually_[9] evidence_[38] evil_[2] examine_[18] example_[32] excellent_[5] exchange_[10] exercise_[13] exist_[25] expense_[7] expose_[2] extend_[12] extreme_[13] fail_[9] fairy_[1] faith_[12] familiar_[8] famous_[9] fan_[7] fashion_[6] favour_[1] feature_[10] female_[14] file_[20] finance_[45] firm_[29] flag_[9] flash_[6] flight_[8] flip_[3] flow_[10] fold_[6] folk_[15] fool_[3] foreign_[41] forgive_[3] forth_[14] fortune_[6] frame_[4] frankly_[5] fruit_[10] fry_[3] fund_[75] furniture_[2] future_[29] gain_[35] garage_[4] gate_[6] gather_[16] gay_[12] gear_[3] generation_[17] gentleman_[4] ghost_[2] gift_[16] glory_[2] goal_[39] golf_[14] grab_[9] grade_[18] grand_[2] grant_[6] grin_[3] grocer_[4] guarantee_[2] guard_[13] guest_[9] guide_[5] guilty_[16] habit_[4] handy_[2] harm_[4] heaven_[3] hedge_[1] height_[5] hesitate_[2] hire_[10] honey_[1] hook_[3] hotel_[10] identify_[38] ignore_[5] ill_[13] illustrate_[8] image_[15] immediate_[14] impress_[6] improve_[42] inch_[11] include_[46] income_[37] increase_[80] incredible_[1] indicate_[43] individual_[26] industry_[45] influence_[21] injure_[18] innocent_[4] inspect_[4] instruct_[20] instrument_[7] intend_[3] intense_[11] intent_[3] interview_[20] introduce_[7] investigate_[22] invite_[9] iron_[1] item_[15] jacket_[8] jam_[1] jeans_[10] journey_[6] juice_[23] junior_[8] justice_[14] knee_[23] knife_[7] knowledge_[24] laboratory_[2] labour_[26] lack_[6] language_[33] lawn_[6] lawyer_[16] league_[19] lean_[24] leap_[1] legal_[37] lend_[4] length_[6] lesson_[16] library_[6] licence_[6] lightly_[1] likely_[35] limit_[29] lion_[4] loan_[18] locate_[5] log_[3] lone_[1] loose_[5] loss_[18] lower_[13] magazine_[25] mail_[14] maintain_[12] male_[18] map_[6] mask_[6] mass_[12] match_[4] mate_[2] material_[29] mathematics_[6] measure_[28] meat_[7] medical_[57] medicine_[15] melt_[9] memory_[5] mental_[22] message_[24] metal_[3] metre_[2] microwave_[1] military_[67] minister_[8] minor_[25] mirror_[5] mission_[4] mix_[13] model_[34] modern_[13] moon_[2] motor_[6] mow_[2] murder_[17] muscle_[3] mystery_[4] nail_[1] narrow_[7] native_[16] nerve_[4] nervous_[13] nest_[2] newspaper_[27] northern_[6] nowhere_[8] nut_[3] oak_[2] object_[4] observe_[2] occasion_[5] occur_[16] official_[52] onion_[33] operate_[37] opinion_[11] opportunity_[38] oppose_[11] opposite_[12] option_[21] ordinary_[6] organize_[55] original_[1] otherwise_[6] oven_[12] owe_[4] pan_[17] partner_[12] pat_[4] path_[11] patient_[35] pattern_[11] pause_[10] peace_[34] pen_[2] pension_[6] per_[1] percent_[67] perform_[59] period_[25] physical_[44] piano_[2] pie_[3] pig_[1] pile_[3] pin_[1] pine_[6] pink_[2]

pipe_[2] pitch_[2] plain_[2] plane_[15] planet_[3] plastic_[16] plate_[6] pleasure_[6] plug_[2] pocket_[18] poem_[4] poet_[2] pole_[4] policy_[82] polish_[1] politics_[132] pollute_[8] pool_[6] popular_[17] population_[44] positive_[40] potato_[7] pour_[15] practical_[3] practise_[36] pray_[3] prefer_[4] pregnant_[10] president_[40] pressure_[30] presume_[1] pretend_[1] prevent_[11] previous_[15] pride_[2] prime_[9] print_[3] privacy_[4] private_[52] process_[62] produce_[20] product_[59] profession_[30] progress_[8] project_[27] property_[18] propose_[6] proud_[7] prove_[17] provide_[136] pump_[4] punish_[2] purchase_[6] purpose_[19] quality_[36] quit_[2] quote_[6] range_[22] rapid_[14] rare_[4] ray_[1] react_[11] recall_[2] receive_[52] recipe_[1] recognize_[12] recommend_[4] recover_[4] reduce_[54] refer_[10] refuse_[7] regard_[4] region_[13] register_[4] regular_[10] relax_[3] release_[13] relief_[9] rely_[9] remain_[26] remark_[4] remind_[2] remove_[10] repair_[2] repeat_[11] replace_[4] represent_[18] require_[31] research_[130] reserve_[10] resist_[3] respect_[15] restaurant_[17] result_[76] retire_[14] rice_[6] rip_[6] risk_[50] rob_[1] role_[36] root_[6] row_[10] royal_[2] rub_[6] ruin_[2] rush_[9] salad_[9] salary_[8] sale_[26] salt_[44] sand_[2] sandwich_[4] satisfy_[3] sauce_[16] scale_[17] scene_[10] score_[38] scratch_[2] scream_[5] screen_[22] screw_[1] search_[14] season_[49] secret_[12] section_[6] seed_[6] seek_[13] select_[10] senior_[44] sentence_[14] separate_[5] series_[15] shade_[2] shadow_[11] sharp_[6] shed_[4] sheet_[13] shelf_[2] shell_[1] shelter_[4] shift_[8] shine_[6] shock_[2] shore_[1] shower_[2] signal_[4] silence_[10] silver_[2] similar_[26] sink_[4] site_[16] ski_[4] skill_[61] skirt_[2] slave_[2] slide_[10] smart_[11] smooth_[6] snap_[5] social_[133] society_[33] soil_[5] soldier_[5] solid_[2] somewhat_[4] sore_[1] soul_[4] soup_[5] southern_[10] spare_[2] species_[19] specific_[27] speech_[12] speed_[18] spell_[2] spin_[6] spirit_[8] split_[2] spray_[2] spread_[9] stable_[6] staff_[18] stairs_[10] stamp_[1] standard_[48] steady_[4] steel_[1] stir_[25] stock_[36] storm_[4] stream_[5] strength_[7] stress_[13] stretch_[8] string_[2] strip_[4] stroke_[3] struggle_[9] style_[7] success_[40] suck_[1] suffer_[15] sugar_[33] super_[2] supply_[16] surface_[9] surround_[2] survive_[13] suspect_[1] swallow_[3] swear_[1] sweep_[3] swing_[12] switch_[7] tale_[3] tank_[11] tap_[1] taxi_[1] technology_[51] teenage_[5] tempt_[1] theatre_[2] therefore_[6] thin_[18] threat_[29] thus_[4] ticket_[11] tide_[4] tiny_[6] tip_[5] title_[10] toe_[2] toilet_[2] tomato_[12] tone_[5] tongue_[2] tool_[11] topic_[15] tough_[20] tour_[9] towel_[2] toy_[2] trace_[4] trade_[46] tradition_[34] traffic_[5] transfer_[6] trial_[24] truck_[14] trunk_[4] tune_[5] twin_[2] typical_[3] union_[15] unit_[10] unite_[4] university_[13] upper_[14] upset_[4] value_[56] various_[16] vary_[33] vegetable_[14] vehicle_[7] version_[7] victim_[12] village_[4] violent_[3] vote_[36] wage_[11] wander_[4] warn_[9] weak_[9] weapon_[26] welcome_[14] western_[23] whip_[4] whistle_[1] wing_[10] wipe_[14] wire_[3] wise_[5] witness_[7] wound_[4] wrap_[10] yell_[2]

BNC-COCA-3,000 Families: [fams 675 : types 791 : tokens 5457]

abandon_[1] abort_[10] abroad_[2] absence_[2] abuse_[31] academy_[26] accommodate_[1] accompany_[2] accomplish_[10] accountable_[1] accurate_[6] achieve_[30] acknowledge_[1] acquire_[8] acre_[2] addict_[2] adequate_[2] adjust_[5] administration_[24] administrator_[6] admission_[2] adolescent_[5] adopt_[12] advocate_[4] affirm_[1] agency_[54] agenda_[4] aggressive_[6] agriculture_[4] aim_[4] aircraft_[2] airline_[8] album_[7] alien_[2] allege_[2] ally_[1] alternative_[11] amend_[6] analyse_[46] analyst_[10] ancient_[2] angle_[4] annual_[24] anticipate_[2] anxiety_[4] appropriate_[12] approve_[10] approximate_[5] archaeology_[1] arise_[8] armed_[7] aspect_[14] assault_[5] assembly_[2] assess_[14] asset_[3] assign_[2] assumption_[4] athlete_[15]

atom_[2] attribute_[5] audience_[6] author_[10] authority_[19] award_[7] ban_[2] bargain_[3] barrier_[6] beam_[2] behave_[4] behaviour_[45] belief_[12] bench_[2] bible_[2] biological_[8] bishop_[1] blend_[3] boost_[2] border_[9] broadcast_[2] budget_[32] burden_[6] bureau_[2] cabinet_[2] campaign_[31] cancer_[40] candidate_[18] capture_[2] carbon_[4] carve_[2] category_[5] catholic_[4] celebrate_[8] cell_[26] ceremony_[2] chairman_[22] characteristic_[3] chart_[1] charter_[2] chemical_[14] cite_[8] civil_[27] civilian_[12] civilise_[2] client_[2] climate_[6] clinic_[16] cluster_[5] coalition_[9] code_[9] coin_[1] collapse_[1] colleague_[3] colony_[2] column_[2] comedy_[2] communicate_[19] compensate_[2] compete_[11] complaint_[2] complex_[16] component_[9] comprehensive_[8] conceive_[1] concept_[15] concert_[2] conclude_[4] conclusion_[10] conduct_[22] confer_[22] confidence_[4] confident_[7] confirm_[3] conflict_[10] confront_[4] congress_[24] consent_[3] consequence_[7] conservative_[12] conserve_[1] considerable_[5] consist_[2] consistent_[5] constitution_[8] construct_[12] consult_[3] consume_[34] consumption_[4] contemporary_[5] content_[4] contest_[2] context_[13] contrast_[3] controversy_[4] convention_[8] convert_[1] convey_[2] convict_[2] cooperate_[6] coordinate_[4] core_[4] corporate_[11] correlate_[8] correspondent_[8] counsel_[18] courage_[2] craft_[2] crew_[4] crisis_[12] criteria_[4] critic_[26] criticism_[2] crop_[6] crucial_[4] cruise_[4] currency_[2] curriculum_[14] cycle_[2] damn_[4] data_[52] debate_[14] decade_[27] declare_[4] decline_[3] defend_[5] deficit_[16] define_[5] democracy_[7] democrat_[59] demonstrate_[9] dense_[6] deposit_[2] deputy_[12] description_[5] destruction_[4] device_[6] devote_[4] differ_[7] digital_[8] dimension_[4] diplomat_[3] disabled_[16] disagree_[5] disaster_[2] disc_[3] discount_[2] discriminate_[2] disorder_[15] dispose_[1] dispute_[6] disrupt_[1] distant_[1] distinct_[2] distinguish_[1] distribute_[2] diverse_[13] division_[2] dna_[2] document_[22] domestic_[20] dominant_[4] donate_[3] dose_[2] draft_[2] drain_[4] drift_[2] drill_[4] eager_[1] eastern_[12] edition_[2] effective_[37] efficient_[14] elaborate_[1] electronic_[6] element_[9] elevate_[2] eliminate_[6] elite_[2] emerge_[6] emergency_[12] emit_[10] emphasis_[8] emphasise_[5] enable_[4] encounter_[4] enforce_[12] enhance_[4] ensure_[4] enterprise_[5] equation_[8] era_[2] error_[6] essay_[4] essential_[5] estimate_[7] ethnic_[26] evaluate_[8] evident_[2] evolve_[1] exceed_[2] exception_[3] exclusive_[5] executive_[38] exhibit_[4] expand_[4] experiment_[4] expert_[18] explicit_[2] explore_[10] extensive_[1] extent_[7] external_[2] extract_[1] facility_[12] factor_[48] factory_[4] faculty_[16] fade_[5] failure_[4] fantasy_[2] federal_[103] fee_[9] fibre_[10] fiction_[7] flee_[1] flesh_[2] flexible_[2] focus_[30] formal_[7] formula_[2] foster_[4] foundation_[4] founded_[1] framework_[5] frequency_[2] frequent_[12] fuel_[20] function_[7] fundamental_[11] funeral_[2] gallery_[2] gang_[6] gap_[5] gaze_[4] gender_[15] gene_[2] generate_[8] genetic_[4] geography_[1] gesture_[4] global_[20] gradual_[2] graduate_[30] grain_[2] grasp_[2] grateful_[2] gravity_[2] grip_[5] gross_[9] guideline_[4] guitar_[2] halt_[1] hazard_[3] heal_[3] heel_[2] heritage_[2] hint_[2] hip_[2] holy_[8] horror_[2] host_[8] household_[7] humour_[4] hypothesis_[4] ideal_[1] immigrant_[6] immune_[2] impact_[19] implement_[6] implicate_[2] importance_[11] impose_[1] impression_[8] incentive_[12] independence_[2] independent_[11] index_[6] inevitable_[1] infant_[4] infect_[3] inflate_[4] ingredient_[4] initiate_[4] insight_[10] install_[4] institution_[25] integrate_[4] intellectual_[2] intelligence_[20] interact_[12] interior_[2] internal_[12] international_[74] interpret_[4] intervene_[10] intimate_[3] invent_[2] invest_[54] jail_[8] jet_[2] joint_[6] journal_[8] jury_[20] justify_[1] label_[2] landscape_[2] launch_[9] layer_[3] leather_[8] lecture_[2] legislate_[15] liberal_[15] liberty_[4] likeness_[2] literal_[5] literary_[4] literature_[13] lobby_[2] magnet_[4] majority_[22] manufacture_[4] margin_[4] mayor_[2] mechanic_[5] media_[18] medium_[14] method_[32] minimum_[2] missile_[11] mixture_[10] mobile_[4] moderate_[1] modest_[1] modify_[6] monitor_[2]

moral_[12] mortal_[6] mortgage_[10] motion_[7] motive_[1] multiple_[11] museum_[4] mutual_[5] naked_[6] negative_[10] neglect_[1] negotiate_[2] net_[6] network_[20] nod_[13] nominate_[8] novel_[5] nuclear_[37] numerous_[4] objective_[4] oblige_[2] obtain_[8] occupation_[1] occupy_[5] ocean_[2] offence_[9] offend_[2] opera_[2] oral_[6] organic_[6] oriented_[4] outcome_[6] overall_[2] overcome_[6] overlook_[2] overwhelm_[2] pace_[5] pale_[10] palm_[4] panel_[8] parliament_[1] participant_[22] participate_[21] passage_[3] passenger_[7] pave_[1] peak_[4] peer_[6] penalty_[2] pepper_[39] perceive_[3] perception_[4] permission_[14] permit_[2] personality_[4] personnel_[4] perspective_[7] persuade_[2] phase_[1] phenomenon_[2] philosophy_[2] phrase_[4] pilot_[14] poll_[25] portion_[8] portrait_[2] pose_[12] potential_[19] poverty_[8] powder_[10] precede_[3] precise_[3] predict_[12] presence_[6] preserve_[1] priest_[1] primary_[23] principal_[4] principle_[6] prior_[7] priority_[9] prize_[2] procedure_[12] proceed_[4] professor_[31] profile_[4] profit_[15] profound_[3] prohibit_[2] prominent_[3] promote_[9] proof_[2] proportion_[4] prosecute_[9] protein_[8] provision_[3] psychiatry_[1] psychology_[19] publication_[1] publish_[26] pursue_[8] puzzle_[3] quantity_[4] racial_[6] radiate_[2] radical_[7] raid_[1] rail_[2] random_[4] rape_[8] ratio_[2] raw_[2] rear_[6] receiver_[2] reflect_[14] reform_[39] refuge_[6] regime_[6] regulate_[20] reject_[2] relative_[23] relevant_[3] religion_[9] religious_[34] remote_[4] reproduce_[6] republic_[25] reputation_[1] request_[3] rescue_[4] resemble_[3] reside_[13] resolution_[12] resolve_[9] resort_[2] resource_[35] respond_[17] response_[14] restrain_[1] retail_[6] reveal_[16] revenue_[21] review_[18] romantic_[2] route_[2] routine_[2] rumour_[2] rural_[12] sacrifice_[2] sample_[24] sanction_[2] satellite_[2] satisfaction_[4] scholar_[2] sculpt_[2] secretary_[12] sector_[9] segment_[2] seize_[2] senate_[19] sensitive_[8] session_[7] severe_[9] shortly_[1] shrug_[7] sigh_[1] significance_[4] significant_[91] silent_[7] slice_[12] slope_[5] software_[13] solution_[12] solve_[14] sophisticated_[4] source_[43] sovereign_[1] spill_[6] squeeze_[4] stain_[2] stake_[3] statistic_[13] status_[11] statute_[2] stem_[5] strategy_[27] strict_[2] structure_[15] studio_[4] subsidy_[2] substance_[7] substantial_[5] subtle_[2] suburb_[4] suicide_[9] sum_[5] summit_[2] supervise_[2] supreme_[7] surgery_[14] survey_[29] sustain_[3] symbol_[2] symptom_[1] tackle_[3] tactic_[2] talent_[6] target_[6] task_[16] technical_[8] technique_[8] temperature_[21] tender_[3] tennis_[8] territory_[3] terror_[16] text_[12] theme_[6] theoretical_[4] theory_[17] therapy_[10] tissue_[2] ton_[1] toss_[2] tournament_[6] trail_[13] transition_[4] transmit_[4] transport_[8] treaty_[4] tremendous_[2] trend_[5] tribe_[3] trigger_[2] troop_[9] ultimate_[7] undergo_[7] uniform_[4] unique_[1] unity_[2] universe_[4] urban_[14] urgent_[2] utility_[10] valid_[2] variety_[13] vast_[5] venture_[7] vessel_[2] veteran_[4] vice_[5] victory_[2] violate_[13] violence_[18] virtual_[4] virus_[2] visible_[8] vision_[7] visual_[5] vital_[2] volume_[2] vulnerable_[6] wealth_[6] weigh_[5] welfare_[17] whisper_[2] withdraw_[2] yield_[4] youth_[8] zone_[5]

BNC-COCA-4,000 Families: [fams 282 : types 293 : tokens 802]

abrupt_[1] absent_[1] accustom_[1] acid_[4] acute_[1] administer_[2] adverse_[2] alike_[2] aluminium_[1] amateur_[1] ambitious_[3] anniversary_[6] arena_[2] array_[2] artificial_[2] astronomy_[1] attorney_[17] automobile_[5] ballot_[5] bankrupt_[5] barrel_[4] baseball_[22] bathe_[1] battered_[1] beneficial_[1] bicycle_[2] blade_[2] blink_[2] boarder_[1] bolt_[2] broker_[1] bulb_[2] bull_[2] bullet_[2] cab_[2] cabin_[2] calorie_[7] campus_[4] candle_[2] canvas_[2] capitalist_[1] casualty_[2] certificate_[1] choir_[1] cholesterol_[18] chronic_[8] chunk_[2] classify_[1] clutch_[1] cognitive_[2] comic_[1] compact_[1] comparative_[1] compel_[4] compile_[1] con_[1] concession_[1]

consensus_[2] consistency_[2] conspiracy_[2] convenience_[2] copyright_[1] cord_[1] corps_[1] couch_[2] coup_[2] crude_[1] cube_[3] custody_[4] debut_[2] deduct_[1] deer_[1] defect_[1] demography_[6] diabetes_[4] diagnose_[2] diagnosis_[3] dial_[2] dignity_[2] dilemma_[2] dim_[2] dioxide_[2] distress_[1] dividend_[1] domain_[2] drown_[2] elbow_[2] elementary_[20] empirical_[3] enact_[2] enrol_[5] equity_[2] ethical_[4] ethics_[2] exert_[2] expertise_[2] explode_[2] eyebrow_[2] fare_[2] fax_[2] fiscal_[4] fist_[1] flour_[22] foresee_[1] fossil_[1] fraction_[4] galaxy_[4] gallon_[1] garlic_[19] geology_[1] glimpse_[4] glove_[3] goat_[2] gravel_[1] greenhouse_[2] habitat_[2] handsome_[2] harass_[2] hardware_[4] haul_[1] helmet_[2] horizontal_[1] hormone_[2] hostage_[5] hug_[1] identical_[4] immigrate_[10] impair_[2] incidence_[1] indigenous_[2] indirect_[2] informal_[2] intact_[2] integral_[1] intelligent_[2] interim_[1] jerk_[1] judicial_[4] kneel_[2] knot_[1] ladder_[3] lap_[4] laser_[2] laundry_[3] leaf_[8] legacy_[2] legislature_[2] leisure_[2] lemon_[11] lens_[2] lesbian_[2] lick_[1] lieutenant_[1] limb_[3] loom_[1] lung_[4] magnitude_[2] medal_[8] metropolitan_[4] militant_[1] milligram_[2] mineral_[2] momentum_[2] monetary_[3] monument_[2] nightmare_[2] noon_[1] norm_[2] notorious_[1] obey_[1] obstacle_[4] obstruct_[1] olive_[7] optimist_[1] ounce_[2] ozone_[1] patrol_[4] peel_[4] pencil_[2] physician_[4] pill_[2] pillow_[2] plaintiff_[2] plea_[4] plead_[2] poke_[2] polar_[2] predominant_[1] preliminary_[2] premium_[4] prescribe_[1] prescription_[2] prestige_[1] prop_[1] questionnaire_[2] rack_[5] recipient_[2] recreation_[1] regain_[2] regress_[4] residue_[2] rib_[2] rifle_[2] roast_[1] rocked_[1] rubber_[2] scarce_[1] scenario_[1] scholarship_[2] scrap_[1] script_[4] secular_[2] seminar_[1] senator_[4] shallow_[2] sheriff_[1] simulate_[1] sip_[2] slam_[4] slap_[4] sleeve_[3] slot_[2] snatch_[1] soak_[3] soap_[2] soccer_[8] socialist_[1] sodium_[7] solar_[10] span_[2] spectacular_[2] spectrum_[4] sphere_[2] spine_[1] spit_[2] splash_[1] spouse_[2] stack_[1] stadium_[2] stance_[2] steep_[2] steer_[1] stimulus_[1] strand_[1] straw_[2] streak_[2] stride_[2] stuffed_[1] surgeon_[4] tablespoon_[43] tag_[2] telescope_[6] terminal_[1] testimony_[2] texture_[2] thumb_[4] tick_[2] tile_[2] tilt_[1] tobacco_[6] tolerate_[3] toxic_[6] trait_[2] transcript_[2] trauma_[1] tremble_[1] tribute_[2] tropics_[4] trustee_[1] tuck_[3] tumble_[1] tumour_[2] utter_[1] vanish_[1] verdict_[2] vertical_[3] virgin_[1] vitamin_[2] wagon_[2] wan_[1] warrant_[1] wrist_[2]

BNC-COCA-5,000 Families: [fams 134 : types 134 : tokens 265]

advocacy_[1] aide_[4] aisle_[6] allergy_[1] aloud_[2] altitude_[1] anecdote_[1] ass_[4] authoritarian_[1] bail_[1] bald_[1] basketball_[32] botany_[1] bulletin_[1] candy_[2] cane_[1] carbohydrate_[13] cardboard_[1] cellular_[1] chord_[1] clap_[1] clasp_[1] cleanse_[1] clench_[1] cocaine_[2] cock_[1] cocktail_[1] coefficient_[1] comb_[1] condom_[2] consecutive_[6] cosmetic_[1] cowboy_[2] crane_[1] cue_[2] dairy_[1] dart_[1] degrade_[1] detention_[1] deviate_[2] dice_[1] diesel_[1] divert_[1] earnest_[1] erosion_[1] facial_[2] fend_[1] foil_[1] ginger_[2] gram_[5] grate_[3] grit_[1] hike_[2] hockey_[4] humanitarian_[1] implant_[1] incur_[1] inmate_[2] intercourse_[1] intrinsic_[1] junk_[2] lash_[1] lightning_[4] lime_[1] liquor_[1] mall_[2] marital_[1] median_[1] medication_[2] memoir_[1] memorable_[1] migrant_[1] millennium_[1] mug_[1] multinational_[2] nominee_[5] oath_[1] oval_[1] paradigm_[1] pharmaceutical_[1] pickup_[2] picnic_[1] porch_[8] pork_[1] pounding_[1] precaution_[1] proximity_[1] purse_[1] qualitative_[1] quantum_[1] radar_[2] ranch_[2] reunion_[1] revolve_[1] rite_[1] saturate_[2] scoop_[1] serial_[1] shorts_[1] shrub_[1] shuttle_[2] sibling_[2] simmer_[4] skate_[1] sneak_[1] sofa_[2] sour_[1] sprinkle_[2] stark_[1] stool_[1] superintendent_[1] surgical_[1] surveillance_[2] sweater_[2] testify_[2] theological_[1] thereafter_[1] toll_[4] transit_[4] trickle_[1]

tub_[1] turtle_[1] undergraduate_[2] uphold_[1] usher_[1] vacation_[7] vacuum_[1] vest_[1] veto_[1] vinegar_[5] void_[1] wheat_[2] whisk_[2] wilderness_[2]

BNC-COCA-6,000 Families: [fams 67 : types 70 : tokens 99]

accord_[1] aerobics_[1] bachelor_[1] barb_[1] bog_[1] booth_[2] calcium_[1] cardiac_[1] churn_[1] closet_[2] clove_[3] colon_[1] columnist_[1] conjure_[1] coronary_[1] crumb_[1] cute_[3] dean_[1] diploma_[2] duct_[1] dune_[1] eldest_[1] embargo_[2] esteem_[2] firefight_[1] fluorescent_[1] focal_[1] freak_[1] freelance_[1] garbage_[2] genome_[1] irrigate_[1] loaf_[1] locker_[1] locus_[1] makeup_[6] martial_[2] mash_[1] mince_[4] mule_[1] multicultural_[1] mustard_[1] nap_[1] pant_[3] parsley_[4] payroll_[1] peanut_[2] pornography_[1] poultry_[1] pre_[1] proliferate_[1] reap_[1] rebound_[2] resonance_[1] retard_[2] slippery_[1] soda_[3] syndicate_[1] tab_[1] trafficked_[2] transcribe_[1] tuition_[2] turbine_[2] vain_[1] vapour_[1] wield_[1] zoom_[1]

BNC-COCA-7,000 Families: [fams 31 : types 31 : tokens 42]

amino_[1] anonymity_[1] anthem_[1] broth_[3] cardiovascular_[1] cinnamon_[2] clockwise_[1] cookie_[5] crank_[1] ethic_[1] flea_[1] guinea_[1] gust_[1] hind_[1] hispanic_[2] homage_[1] illicit_[1] longitudinal_[1] margarine_[1] marijuana_[2] marrow_[1] motel_[1] outstretched_[1] prostate_[1] punitive_[1] socioeconomic_[1] teddy_[1] transfuse_[1] trooper_[1] vanilla_[3] vantage_[1]

BNC-COCA-8,000 Families: [fams 23 : types 23 : tokens 29]

aback_[1] arsenal_[1] ballistic_[1] basil_[1] blurt_[1] concerted_[1] confection_[1] disobedient_[1] embryonic_[2] freshman_[4] gasoline_[2] gobble_[1] hone_[1] jot_[1] metro_[1] monoxide_[1] ovary_[1] parochial_[1] rookie_[1] soy_[1] thyme_[2] yolk_[1] zest_[1]

BNC-COCA-9,000 Families: [fams 12 : types 12 : tokens 22]

appellate_[1] denominator_[1] globule_[1] granulate_[2] headway_[1] herein_[5] multivariate_[1] nutmeg_[1] ovation_[1] playoff_[6] sclerosis_[1] wishful_[1]

BNC-COCA-10,000 Families: [fams 5 : types 5 : tokens 5]

emeritus_[1] grizzly_[1] kosher_[1] prenatal_[1] supra_[1]

BNC-COCA-11,000 Families: [fams 6 : types 6 : tokens 10]

boomer_[1] cayenne_[1] cumin_[2] duffel_[1] quo_[1] skillet_[4]

BNC-COCA-12,000 Families: [fams 2 : types 2 : tokens 2]

capita_[1] extracurricular_[1]

BNC-COCA-13,000 Families: [fams 1 : types 1 : tokens 1]

litmus_[1]

BNC-COCA-14,000 Families: [fams 3 : types 3 : tokens 3]

cilantro_[1] deco_[1] sophomore_[1]

BNC-COCA-15,000 Families: [fams : types : tokens]

BNC-COCA-16,000 Families: [fams : types : tokens]

BNC-COCA-17,000 Families: [fams 2 : types 2 : tokens 2]

canola_[1] neutron_[1]

BNC-COCA-18,000 Families: [fams : types : tokens]

BNC-COCA-19,000 Families: [fams : types : tokens]

BNC-COCA-20,000 Families: [fams : types : tokens]

BNC-COCA-21,000 Families: [fams : types : tokens]

BNC-COCA-22,000 Families: [fams : types : tokens]

BNC-COCA-23,000 Families: [fams : types : tokens]

BNC-COCA-24,000 Families: [fams : types : tokens]

BNC-COCA-25,000 Families: [fams : types : tokens]

OFFLIST: [?: types 87 : tokens 365]

african_[10] airplane_[2] airport_[5] american_[47] arab_[5] asian_[3] bathroom_[3] bedroom_[10] birthday_[6] bitch_[1] british_[4] canadian_[1] ceo_[1] chinese_[2] christian_[6] classroom_[17] comeback_[1] doorbell_[1] doorway_[2] download_[1] downstairs_[1] downturn_[1] driveway_[2] dropout_[3] dutch_[1] english_[8] european_[6] feedback_[4] footstep_[2] forever_[13] french_[4] fuck_[1] groundwork_[1] halfway_[2] hallway_[2] hardwood_[1] headline_[2] homeland_[2] homer_[1] horseback_[1] indian_[4] iraqi_[10] islamic_[3] israeli_[7] japanese_[2] jewish_[2] laptop_[1] latin_[1] lawsuit_[2] lifestyle_[2] lineman_[1] longtime_[2] mainstream_[2] mexican_[2] mph_[1] muslim_[4] nongovernmental_[1] olympic_[4] olympics_[1] online_[4] palestinian_[3] pc_[1] persian_[1] piss_[1] pm_[5] preservice_[1] proofread_[1] railroad_[2] russian_[1] saucepan_[4] shit_[3] someday_[2] someplace_[1] southeast_[1] soviet_[6] spanish_[1] spokesman_[2] spokeswoman_[1] subscale_[1] teaspoon_[62] thanksgiving_[1] touchdown_[8] upstairs

Appendix 5. Cut-off of one occurrence per 500,000 tokens resulting in 4,778 word families plus off-list types

WORD LIST	TOKENS/%	TYPES/%	FAMILIES
1	64937/56.40	1546/25.29	908
2	28281/24.57	1382/22.61	916
3	15864/13.78	1220/19.96	913
4	2950/ 2.56	700/11.45	639
5	908/ 0.79	395/ 6.46	382
6	342/ 0.30	203/ 3.32	197
7	193/ 0.17	132/ 2.16	129
8	124/ 0.11	80/ 1.31	80
9	73/ 0.06	44/ 0.72	43
10	43/ 0.04	37/ 0.61	36
11	35/ 0.03	22/ 0.36	22
12	13/ 0.01	12/ 0.20	12
13	7/ 0.01	7/ 0.11	7
14	14/ 0.01	7/ 0.11	7
15	3/ 0.00	3/ 0.05	3
16	2/ 0.00	2/ 0.03	2
17	4/ 0.00	4/ 0.07	4
18	0/ 0.00	0/ 0.00	0
19	0/ 0.00	0/ 0.00	0
20	0/ 0.00	0/ 0.00	0
21	0/ 0.00	0/ 0.00	0
22	0/ 0.00	0/ 0.00	0
23	1/ 0.00	1/ 0.02	1
24	1/ 0.00	1/ 0.02	1
25	0/ 0.00	0/ 0.00	0
26	0/ 0.00	0/ 0.00	0
27	0/ 0.00	0/ 0.00	0
28	0/ 0.00	0/ 0.00	0
29	0/ 0.00	0/ 0.00	0
30	0/ 0.00	0/ 0.00	0
31	378/ 0.33	41/ 0.67	40
32	21/ 0.02	7/ 0.11	6
33	605/ 0.53	118/ 1.93	116
34	22/ 0.02	8/ 0.13	8
Not in the lists	306/ 0.27	140/ 2.29	?????
Total	115127	6112	4472

Appendix 6. 12,615 lemma pairs remaining after duplicates such as take-walk/ walk-take were removed from 25,969 lemma pairs
(This is just a sample. Full and more detailed data available upon request)

PIVOT WORD	PART OF SPEECH	COLLOCATE	PART OF SPEECH	FREQUENCY
be	v	think	v	371805
be	v	now	r	293971
be	v	here	r	247641
be	v	very	r	237126
be	v	more	r	222430
be	v	how	r	221721
do	v	know	v	218547
be	v	there	r	200045
be	v	thing	n	192459
be	v	good	j	182608
be	v	year	n	182208
be	v	also	r	179357
be	v	way	n	165443
be	v	still	r	161143
do	v	think	v	156252
be	v	really	r	148749
be	v	too	r	146393
do	v	want	v	139693
be	v	only	r	137429
be	v	why	r	125564
be	v	where	r	121899
do	v	how	r	111075
do	v	how	r	111075
be	v	important	j	104366
be	v	sure	j	102842
be	v	always	r	101910
be	v	most	r	100082
be	v	great	j	95838
be	v	big	j	94338
be	v	right	j	94041
have	v	never	r	92868
be	v	part	n	92686
be	v	problem	n	91806
have	v	year	n	88481
be	v	question	n	87034
year	n	ago	r	84034
be	v	about	r	82660
will	v	have	v	81308
be	v	happen	v	76577
be	v	fact	n	74035

be	v	believe	v	73910
be	v	today	r	72414
be	v	different	j	71874
be	v	bad	j	71794
be	v	place	n	70969
be	v	only	j	70634
be	v	case	n	70299
do	v	why	r	67581
be	v	enough	r	66708
be	v	likely	j	64844
be	v	TRUE	j	64512
right	r	all	r	62559
be	v	point	n	60230
as	r	well	r	58717
be	v	hard	j	58672
be	v	real	j	58442
be	v	issue	n	57937
be	v	much	r	57393
be	v	already	r	57040
be	v	reason	n	55215
be	v	work	n	55025
be	v	best	j	54883
know	v	how	r	54656
be	v	far	r	54474
be	v	often	r	54383
school	n	high	j	54032
be	v	story	n	52641
be	v	kind	n	51969
be	v	consider	v	51568
be	v	easy	j	51386
up	r	pick	v	50745
be	v	probably	r	50719
up	r	come	v	50603
already	r	have	v	50331
go	v	back	r	50212
come	v	back	r	49953
be	v	actually	r	49335
be	v	expect	v	49278
have	v	ever	r	48848
now	r	right	r	48480
be	v	yet	r	48428
be	v	system	n	48210
be	v	almost	r	47983
be	v	wrong	j	47549
be	v	person	n	47444
see	v	can	v	47299
be	v	however	r	47063

do	v	mean	v	46988
will	v	say	v	46349
go	v	out	r	46326
be	v	difficult	j	46325
do	v	really	r	46268
be	v	idea	n	45767
be	v	business	n	45555
be	v	government	n	45340
can	v	get	v	44999
have	v	hear	v	44461
be	v	clear	j	44354
be	v	supposed	j	43837
be	v	possible	j	43145

Appendix 7. Items flagged at the 2.5 percent parameter and native speaker judgments
(This is just a sample. Full and more detailed data available upon request)

NATIVE SPEAKER JUDGMENT 1-5	PIVOT WORD	PART OF SPEECH	COLLOCATE	PART OF SPEECH
1	note	n	supra	n
1	teacher	n	preservice	n
1	control	n	locus	n
1	set	v	current	j
1	set	v	result	n
1	current	j	search	n
1	cluster	n	globular	j
1	spring	n	sandy	j
1	democrat	n	representative	j
1	nation	n	talk	n
2	capital	n	gang	n
2	standard	j	deviation	n
2	service	n	reader	n
2	variable	n	dependent	j
2	factor	n	analysis	n
2	independent	j	variable	n
2	analysis	n	regression	n
2	service	n	card	n
2	reader	n	circle	n
2	service	n	circle	n
2	internal	j	consistency	n
3	area	n	content	j
3	study	n	present	j
3	visual	j	impairment	n
3	significant	j	statistically	r
3	social	j	support	n
3	data	n	collection	n
3	analysis	n	use	v
3	difference	n	gender	n
3	school	n	psychologist	n
3	music	n	educator	n
4	study	n	social	j
4	school	n	counselor	n
4	education	n	music	n
4	study	n	examine	v
4	effect	n	significant	j
4	study	n	purpose	n
4	result	n	indicate	v
4	research	n	future	j
4	difference	n	group	n
4	student	n	gifted	j
5	education	n	physical	j
5	difference	n	significant	j
5	knowledge	n	student	n

5	level	n	report	v
5	study	n	future	j
5	learning	n	teaching	n
5	short	j	while	n
5	difference	n	examine	v
5	level	n	significantly	r
5	fire	n	firefighter	n
5	right	r	all	r

Appendix 8. Items flagged at the 5 percent parameter and native speaker judgments
(This is just a sample. Full and more detailed data available upon request)

NATIVE SPEAKER JUDGMENT 1-5	PIVOT WORD	PART OF SPEECH	COLLOCATE	PART OF SPEECH
1	result	n	search	n
1	african	j	student	n
1	president	n	marketing	n
1	pc	n	world	n
1	say	v	executive	n
1	conference	n	western	j
1	pm	r	eastern	j
1	chief	j	correspondent	n
1	note	n	supra	n
1	teacher	n	preservice	n
1	control	n	locus	n
2	card	n	reader	n
2	social	j	structure	n
2	status	n	socioeconomic	j
2	soft	j	tissue	n
2	matter	n	organic	j
2	cultural	j	context	n
2	democracy	n	liberal	j
2	social	j	psychological	j
2	middle	j	ear	n
2	model	n	test	v
2	social	j	psychology	n
3	work	n	social	j
3	education	n	special	j
3	study	n	result	n
3	data	n	analysis	n
3	teacher	n	classroom	n
3	control	n	group	n
3	study	n	finding	n
3	program	n	teacher	n
3	student	n	experience	n
3	social	j	skill	n
4	community	n	college	n
4	student	n	teacher	n
4	activity	n	physical	j
4	male	n	female	n
4	skill	n	knowledge	n
4	find	v	difference	n
4	high	j	significantly	r
4	find	v	significant	j
4	student	n	skill	n
4	opportunity	n	student	n
5	useful	j	may	v
5	eye	n	close	v
5	ago	r	hour	n

5	walk	v	slowly	r
5	face	n	expression	n
5	door	n	open	r
5	man	n	stare	v
5	door	n	shut	j
5	face	n	wash	v
5	blow	v	nose	n

Appendix 9. Items flagged at the 10 percent parameter and native speaker judgments
(This is just a sample. Full and more detailed data available upon request)

NATIVE SPEAKER JUDGMENT 1-5	PIVOT WORD	PART OF SPEECH	COLLOCATE	PART OF SPEECH
1	result	n	search	n
1	african	j	student	n
1	president	n	marketing	n
1	pc	n	world	n
1	say	v	executive	n
1	conference	n	western	j
1	pm	r	eastern	j
1	chief	j	correspondent	n
1	note	n	supra	n
1	teacher	n	preservice	n
1	control	n	locus	n
2	card	n	reader	n
2	social	j	structure	n
2	status	n	socioeconomic	j
2	soft	j	tissue	n
2	matter	n	organic	j
2	cultural	j	context	n
2	democracy	n	liberal	j
2	social	j	psychological	j
2	middle	j	ear	n
2	model	n	test	v
3	work	n	social	j
3	education	n	special	j
3	study	n	result	n
3	data	n	analysis	n
3	teacher	n	classroom	n
3	control	n	group	n
3	study	n	finding	n
3	program	n	teacher	n
3	student	n	experience	n
3	social	j	skill	n
4	community	n	college	n
4	student	n	teacher	n
4	activity	n	physical	j
4	male	n	female	n
4	skill	n	knowledge	n
4	find	v	difference	n
4	high	j	significantly	r
4	find	v	significant	j
4	student	n	skill	n
4	opportunity	n	student	n
5	useful	j	may	v
5	eye	n	close	v
5	ago	r	hour	n

5	walk	v	slowly	r
5	face	n	expression	n
5	door	n	open	r
5	man	n	stare	v
5	door	n	shut	j
5	face	n	wash	v
5	blow	v	nose	n

Appendix 10. Not flagged but judged to have issues at 2.5 percent 2,088
(This is just a sample. Full and more detailed data available upon request)

TYPE	PIVOT WORD	PART OF SPEECH	COLLOCATE	PART OF SPEECH
body	smile	n	eye	n
body	shorts	n	wear	v
body	shirt	n	jeans	n
body	run	v	foot	n
body	roll	v	head	n
chronological	word	n	processor	n
chronological	video	n	digital	j
chronological	sell	v	record	n
chronological	room	n	sitting	n
chronological	new	j	millennium	n
color	yellow	j	black	j
color	white	j	yellow	j
color	white	j	wine	n
color	white	j	wear	v
color	white	j	wall	n
direction	western	j	eastern	j
direction	western	j	art	n
direction	west	n	wing	n
direction	west	n	east	n
direction	west	n	coast	n
fiction	winter	n	night	n
fiction	window	n	room	n
fiction	window	n	picture	n
fiction	window	n	glass	n
fiction	window	n	front	j
noise	weekend	n	day	n
noise	water	n	plant	n
noise	wall	n	hole	n
noise	tree	n	forest	n
noise	score	n	sit	v
not useful	year	n	decade	n
not useful	would	v	dollar	n
not useful	wheel	n	spin	v
not useful	wall	n	window	n
not useful	wall	n	room	n
proper noun	world	n	trade	n
proper noun	world	n	series	n
proper noun	world	n	cup	n
proper noun	world	n	bank	n
proper noun	western	j	world	n
specialized	tape	n	play	v
specialized	yard	n	run	v
specialized	worker	n	care	n
specialized	word	n	letter	n

specialized woman n battered j

Appendix 11. Not flagged but judged to have issues at 5 percent 1,788
(This is just a sample. Full data available upon request)

TYPE	PIVOT WORD	PART OF SPEECH	COLLOCATE	PART OF SPEECH
body	right	j	shoulder	n
body	reach	v	foot	n
body	put	v	foot	n
body	out	r	hand	v
body	neck	n	down	r
chronological	word	n	processor	n
chronological	video	n	digital	j
chronological	sell	v	record	n
chronological	room	n	sitting	n
chronological	new	j	millennium	n
color	white	j	suit	n
color	white	j	shirt	n
color	white	j	sheet	n
color	white	j	red	n
color	white	j	red	j
direction	west	n	central	j
direction	turn	v	left	r
direction	town	n	southern	j
direction	state	n	western	j
direction	state	n	southern	j
fiction	wear	v	sweater	n
fiction	wear	v	skirt	n
fiction	wear	v	shoe	n
fiction	wear	v	shirt	n
fiction	wear	v	red	j
noise	other	j	foot	n
noise	music	n	band	n
noise	information	n	please	r
noise	help	v	similar	j
noise	far	j	please	r
not useful	wall	n	line	v
not useful	wall	n	glass	n
not useful	wall	n	floor	n
not useful	wall	n	cover	v
not useful	wall	n	ceiling	n
proper noun	western	j	tradition	n
proper noun	western	j	nation	n
proper noun	war	n	cold	j
proper noun	vehicle	n	utility	n
proper noun	university	n	state	n
specialized	win	v	super	j
specialized	win	v	race	n
specialized	win	v	medal	n
specialized	win	v	gold	n

specialized	whole	j	wheat	n

Appendix 12. Not flagged but judged to have issues at 10 percent 1,193
(This is just a sample. Full data available upon request)

TYPE	PIVOT WORD	PART OF SPEECH	COLLOCATE	PART OF SPEECH
body	control	n	arm	n
body	ball	n	foot	n
body	back	n	foot	n
body	arm	n	lift	v
body	arm	n	left	j
chronological	word	n	processor	n
chronological	video	n	digital	j
chronological	sell	v	record	n
chronological	room	n	sitting	n
chronological	new	j	millennium	n
color	black	j	coat	n
color	black	j	brown	j
color	black	j	boot	n
color	black	j	bear	n
color	black	j	bag	n
direction	country	n	eastern	j
direction	come	v	foot	n
direction	city	n	northern	j
direction	central	j	eastern	j
direction	beach	n	down	r
fiction	door	n	sliding	j
fiction	door	n	double	j
fiction	chair	n	room	n
fiction	bag	n	hold	v
fiction	air	n	hang	v
noise	contain	v	quote	n
noise	computer	n	use	v
noise	change	n	undergo	v
noise	attractive	j	make	v
noise	approach	n	more	r
not useful	bed	n	sleep	v
not useful	bag	n	large	j
not useful	art	n	music	n
not useful	art	n	artist	n
not useful	air	n	hot	j
proper noun	administration	n	safety	n
proper noun	administration	n	national	j
proper noun	administration	n	health	n
proper noun	ad	n	agency	n
proper noun	action	n	affirmative	j
specialized	tape	n	play	v
specialized	am	r	morning	n
specialized	am	r	around	r
specialized	aide	n	say	v

specialized	agency	n	management	n
specialized	age	n	year	n

Appendix 14. Items flagged at 2.5 for chronological issues and native speaker judgments
(More detailed data available upon request)

JUDGMENT BY NATIVE	PIVOT WORD	PART OF SPEECH	COLLOCATE	PART OF SPEECH
Accurately Flagged	budget	n	amendment	n
Accurately Flagged	amendment	n	balanced	j
Accurately Flagged	suicide	n	bomber	n
Accurately Flagged	marriage	n	gay	j
Accurately Flagged	research	n	cell	n
Accurately Flagged	marriage	n	same-sex	j
Accurately Flagged	cell	n	embryonic	j
Accurately Flagged	stem	n	embryonic	j
Accurately Flagged	care	n	managed	j
Accurately Flagged	bill	n	crime	n
Accurately Flagged	package	n	stimulus	n
Accurately Flagged	research	n	stem	n
Accurately Flagged	suicide	n	bombing	n
Accurately Flagged	government	n	interim	j
Accurately Flagged	cell	n	stem	n
Inaccurately Flagged	low	j	vision	n
Inaccurately Flagged	school	n	counseling	n
Inaccurately Flagged	report	v	participant	n
Inaccurately Flagged	represent	v	text	n
Inaccurately Flagged	represent	v	equation	n
Inaccurately Flagged	line	n	equation	n
Inaccurately Flagged	limb	n	residual	j
Inaccurately Flagged	nerve	n	facial	j
Inaccurately Flagged	food	n	residuals	n
Inaccurately Flagged	set	v	result	n
Inaccurately Flagged	set	v	current	j
Inaccurately Flagged	current	j	search	n
Inaccurately Flagged	result	n	search	n
Inaccurately Flagged	sat	v	fat	n
Inaccurately Flagged	play	v	football	n
Inaccurately Flagged	winner	n	match	n
Inaccurately Flagged	video	n	clip	n
Inaccurately Flagged	begin	v	clip	n
Inaccurately Flagged	video	n	begin	v
Inaccurately Flagged	moment	n	break	n
Inaccurately Flagged	moment	n	commercial	j
Inaccurately Flagged	speak	v	interpreter	n
Inaccurately Flagged	continue	v	prime-time	n
Inaccurately Flagged	site	n	web	n
Inaccurately Flagged	phone	n	cell	n
Inaccurately Flagged	send	v	e-mail	n
Inaccurately Flagged	war	n	terror	n
Inaccurately Flagged	war	n	terrorism	n

Inaccurately Flagged	e-mail	n	address	n
Inaccurately Flagged	fiber	n	carbohydrate	n
Inaccurately Flagged	internet	n	use	v
Inaccurately Flagged	camera	n	digital	j
Inaccurately Flagged	internet	n	service	n
Inaccurately Flagged	internet	n	site	n
Inaccurately Flagged	search	n	engine	n
Inaccurately Flagged	get	v	e-mail	n
Inaccurately Flagged	show	n	reality	n
Inaccurately Flagged	visit	v	information	n
Inaccurately Flagged	phone	n	e-mail	n
Inaccurately Flagged	speed	n	mixer	n
Inaccurately Flagged	go	v	online	r
Inaccurately Flagged	room	n	chat	n
Inaccurately Flagged	check	v	site	n
Inaccurately Flagged	disorder	n	bipolar	j
Inaccurately Flagged	people	n	when	r
Inaccurately Flagged	independent	j	counsel	n
Inaccurately Flagged	serve	v	purpose	n
Inaccurately Flagged	best	j	player	n
Inaccurately Flagged	internet	n	company	n
Inaccurately Flagged	minute	n	preparation	n
Inaccurately Flagged	fat	n	sat	n
Inaccurately Flagged	seem	v	obvious	j
Inaccurately Flagged	attack	n	terror	n
Inaccurately Flagged	fight	v	terrorism	n
Inaccurately Flagged	e-mail	n	fax	v
Inaccurately Flagged	preparation	n	serving	n
Inaccurately Flagged	back	r	rock	v
Inaccurately Flagged	add	v	additional	j

Appendix 14. Items flagged at 2.5 for chronological issues and native speaker judgments
(More detailed data available upon request)

JUDGMENT BY NATIVE	PIVOT WORD	PART OF SPEECH	COLLOCATE	PART OF SPEECH
Accurately Flagged	budget	n	amendment	n
Accurately Flagged	amendment	n	balanced	j
Accurately Flagged	suicide	n	bomber	n
Accurately Flagged	marriage	n	gay	j
Accurately Flagged	research	n	cell	n
Accurately Flagged	marriage	n	same-sex	j
Accurately Flagged	cell	n	embryonic	j
Accurately Flagged	stem	n	embryonic	j
Accurately Flagged	care	n	managed	j
Accurately Flagged	bill	n	crime	n
Accurately Flagged	package	n	stimulus	n
Accurately Flagged	research	n	stem	n
Accurately Flagged	suicide	n	bombing	n
Accurately Flagged	government	n	interim	j
Accurately Flagged	cell	n	stem	n
Inaccurately Flagged	low	j	vision	n
Inaccurately Flagged	school	n	counseling	n
Inaccurately Flagged	report	v	participant	n
Inaccurately Flagged	represent	v	text	n
Inaccurately Flagged	represent	v	equation	n
Inaccurately Flagged	line	n	equation	n
Inaccurately Flagged	limb	n	residual	j
Inaccurately Flagged	nerve	n	facial	j
Inaccurately Flagged	food	n	residuals	n
Inaccurately Flagged	set	v	result	n
Inaccurately Flagged	set	v	current	j
Inaccurately Flagged	current	j	search	n
Inaccurately Flagged	result	n	search	n
Inaccurately Flagged	sat	v	fat	n
Inaccurately Flagged	play	v	football	n
Inaccurately Flagged	winner	n	match	n
Inaccurately Flagged	video	n	clip	n
Inaccurately Flagged	begin	v	clip	n
Inaccurately Flagged	video	n	begin	v
Inaccurately Flagged	moment	n	break	n
Inaccurately Flagged	moment	n	commercial	j
Inaccurately Flagged	speak	v	interpreter	n
Inaccurately Flagged	continue	v	prime-time	n
Inaccurately Flagged	site	n	web	n
Inaccurately Flagged	phone	n	cell	n
Inaccurately Flagged	send	v	e-mail	n
Inaccurately Flagged	war	n	terror	n
Inaccurately Flagged	war	n	terrorism	n

Inaccurately Flagged	e-mail	n	address	n
Inaccurately Flagged	fiber	n	carbohydrate	n
Inaccurately Flagged	internet	n	use	v
Inaccurately Flagged	camera	n	digital	j
Inaccurately Flagged	internet	n	service	n
Inaccurately Flagged	internet	n	site	n
Inaccurately Flagged	search	n	engine	n
Inaccurately Flagged	get	v	e-mail	n
Inaccurately Flagged	show	n	reality	n
Inaccurately Flagged	visit	v	information	n
Inaccurately Flagged	phone	n	e-mail	n
Inaccurately Flagged	speed	n	mixer	n
Inaccurately Flagged	go	v	online	r
Inaccurately Flagged	room	n	chat	n
Inaccurately Flagged	check	v	site	n
Inaccurately Flagged	disorder	n	bipolar	j
Inaccurately Flagged	people	n	when	r
Inaccurately Flagged	independent	j	counsel	n
Inaccurately Flagged	serve	v	purpose	n
Inaccurately Flagged	best	j	player	n
Inaccurately Flagged	internet	n	company	n
Inaccurately Flagged	minute	n	preparation	n
Inaccurately Flagged	fat	n	sat	n
Inaccurately Flagged	seem	v	obvious	j
Inaccurately Flagged	attack	n	terror	n
Inaccurately Flagged	fight	v	terrorism	n
Inaccurately Flagged	e-mail	n	fax	v
Inaccurately Flagged	preparation	n	serving	n
Inaccurately Flagged	back	r	rock	v
Inaccurately Flagged	add	v	additional	j

Appendix 15. Items flagged at 5 percent for chronological issues and native speaker judgments
(More detailed data available upon request)

JUDGMENT BY NATIVE	PIVOT WORD	PART OF SPEECH	COLLOCATE	PART OF SPEECH
Accurately Flagged	budget	n	amendment	n
Accurately Flagged	amendment	n	balanced	j
Accurately Flagged	suicide	n	bomber	n
Accurately Flagged	marriage	n	gay	j
Accurately Flagged	research	n	cell	n
Accurately Flagged	marriage	n	same-sex	j
Accurately Flagged	cell	n	embryonic	j
Accurately Flagged	stem	n	embryonic	j
Accurately Flagged	care	n	managed	j
Accurately Flagged	bill	n	crime	n
Accurately Flagged	package	n	stimulus	n
Accurately Flagged	research	n	stem	n
Accurately Flagged	suicide	n	bombing	n
Accurately Flagged	government	n	interim	j
Accurately Flagged	cell	n	stem	n
Accurately Flagged	fund	n	hedge	n
Accurately Flagged	health	n	reform	n
Accurately Flagged	budget	n	balanced	j
Accurately Flagged	care	n	reform	n
Accurately Flagged	party	n	reform	n
Accurately Flagged	force	n	coalition	n
Accurately Flagged	reduction	n	deficit	n
Accurately Flagged	new	j	millennium	n
Accurately Flagged	weapon	n	inspector	n
Accurately Flagged	bond	n	junk	n
Accurately Flagged	tax	n	flat	j
Accurately Flagged	lift	v	embargo	n
Accurately Flagged	reform	n	health-care	n
Inaccurately Flagged	low	j	vision	n
Inaccurately Flagged	school	n	counseling	n
Inaccurately Flagged	report	v	participant	n
Inaccurately Flagged	represent	v	text	n
Inaccurately Flagged	represent	v	equation	n
Inaccurately Flagged	line	n	equation	n
Inaccurately Flagged	limb	n	residual	j
Inaccurately Flagged	nerve	n	facial	j
Inaccurately Flagged	food	n	residuals	n
Inaccurately Flagged	set	v	result	n
Inaccurately Flagged	set	v	current	j
Inaccurately Flagged	current	j	search	n
Inaccurately Flagged	result	n	search	n
Inaccurately Flagged	sat	v	fat	n
Inaccurately Flagged	play	v	football	n

Inaccurately Flagged	winner	n	match	n
Inaccurately Flagged	video	n	clip	n
Inaccurately Flagged	begin	v	clip	n
Inaccurately Flagged	video	n	begin	v
Inaccurately Flagged	moment	n	break	n
Inaccurately Flagged	moment	n	commercial	j
Inaccurately Flagged	speak	v	interpreter	n
Inaccurately Flagged	continue	v	prime-time	n
Inaccurately Flagged	site	n	web	n
Inaccurately Flagged	phone	n	cell	n
Inaccurately Flagged	send	v	e-mail	n
Inaccurately Flagged	war	n	terror	n
Inaccurately Flagged	war	n	terrorism	n
Inaccurately Flagged	e-mail	n	address	n
Inaccurately Flagged	fiber	n	carbohydrate	n
Inaccurately Flagged	internet	n	use	v
Inaccurately Flagged	camera	n	digital	j
Inaccurately Flagged	internet	n	service	n
Inaccurately Flagged	internet	n	site	n
Inaccurately Flagged	search	n	engine	n
Inaccurately Flagged	get	v	e-mail	n
Inaccurately Flagged	show	n	reality	n
Inaccurately Flagged	visit	v	information	n
Inaccurately Flagged	phone	n	e-mail	n
Inaccurately Flagged	speed	n	mixer	n
Inaccurately Flagged	go	v	online	r
Inaccurately Flagged	room	n	chat	n
Inaccurately Flagged	check	v	site	n
Inaccurately Flagged	disorder	n	bipolar	j
Inaccurately Flagged	people	n	when	r
Inaccurately Flagged	independent	j	counsel	n
Inaccurately Flagged	serve	v	purpose	n
Inaccurately Flagged	best	j	player	n
Inaccurately Flagged	internet	n	company	n
Inaccurately Flagged	minute	n	preparation	n
Inaccurately Flagged	fat	n	sat	n
Inaccurately Flagged	seem	v	obvious	j
Inaccurately Flagged	attack	n	terror	n
Inaccurately Flagged	fight	v	terrorism	n
Inaccurately Flagged	e-mail	n	fax	v
Inaccurately Flagged	preparation	n	serving	n
Inaccurately Flagged	back	r	rock	v
Inaccurately Flagged	add	v	additional	j
Inaccurately Flagged	school	n	counselor	n
Inaccurately Flagged	visual	j	impairment	n
Inaccurately Flagged	result	n	current	j
Inaccurately Flagged	method	n	participant	n
Inaccurately Flagged	participant	n	complete	v
Inaccurately Flagged	student	n	impairment	n
Inaccurately Flagged	middle	j	ear	n

Inaccurately Flagged	indicate	v	participant	n
Inaccurately Flagged	symptom	n	depressive	n
Inaccurately Flagged	capital	n	social	j
Inaccurately Flagged	response	n	participant	n
Inaccurately Flagged	note	n	supra	n
Inaccurately Flagged	use	n	substance	n
Inaccurately Flagged	school	n	psychology	n
Inaccurately Flagged	equation	n	can	v
Inaccurately Flagged	note	n	text	n
Inaccurately Flagged	text	n	accompanying	j
Inaccurately Flagged	multiple	j	equation	n
Inaccurately Flagged	note	n	accompanying	j
Inaccurately Flagged	art	n	educator	n
Inaccurately Flagged	environmental	j	knowledge	n
Inaccurately Flagged	fat	n	protein	n
Inaccurately Flagged	cholesterol	n	carbohydrate	n
Inaccurately Flagged	lemon	n	zest	n
Inaccurately Flagged	total	j	minute	n
Inaccurately Flagged	salt	n	kosher	j
Inaccurately Flagged	beat	v	speed	n
Inaccurately Flagged	cook	v	spray	n
Inaccurately Flagged	article	n	copyright	n
Inaccurately Flagged	face	v	page	n
Inaccurately Flagged	fire	n	firefighter	n
Inaccurately Flagged	include	v	survivor	n
Inaccurately Flagged	scene	n	violence	n
Inaccurately Flagged	wife	n	survivor	n
Inaccurately Flagged	game	n	winner	n
Inaccurately Flagged	join	v	conversation	n
Inaccurately Flagged	join	v	phone	n
Inaccurately Flagged	report	n	tonight	r
Inaccurately Flagged	break	n	away	r
Inaccurately Flagged	commercial	j	away	r
Inaccurately Flagged	continue	v	commercial	j
Inaccurately Flagged	capital	n	gang	n
Inaccurately Flagged	attack	n	terrorist	j
Inaccurately Flagged	total	j	fat	n
Inaccurately Flagged	internet	n	access	n
Inaccurately Flagged	sheet	n	baking	n
Inaccurately Flagged	sugar	n	fiber	n
Inaccurately Flagged	message	n	e-mail	n
Inaccurately Flagged	type	n	diabetes	n
Inaccurately Flagged	phone	n	mobile	j
Inaccurately Flagged	can	v	e-mail	n
Inaccurately Flagged	information	n	site	n
Inaccurately Flagged	often	r	stir	v
Inaccurately Flagged	teaspoon	n	olive	j
Inaccurately Flagged	fiber	n	cholesterol	n
Inaccurately Flagged	food	n	organic	j
Inaccurately Flagged	message	n	text	n

Inaccurately Flagged	total	j	carbohydrate	n
Inaccurately Flagged	can	v	download	v
Inaccurately Flagged	can	v	online	r
Inaccurately Flagged	serving	n	minute	n
Inaccurately Flagged	will	v	end	v
Inaccurately Flagged	phone	n	cellular	j
Inaccurately Flagged	break	v	record	n
Inaccurately Flagged	system	n	expert	n
Inaccurately Flagged	welfare	n	recipient	n
Inaccurately Flagged	policy	n	industrial	j
Inaccurately Flagged	money	n	soft	j
Inaccurately Flagged	local	j	regional	j
Inaccurately Flagged	college	n	electoral	j
Inaccurately Flagged	low	j	lip	n
Inaccurately Flagged	scene	n	drug	n
Inaccurately Flagged	internet	n	connection	n
Inaccurately Flagged	insurance	n	deposit	n
Inaccurately Flagged	enter	v	week	n
Inaccurately Flagged	organize	v	help	v
Inaccurately Flagged	threat	n	terrorist	j

Appendix 16. Items flagged at 10 percent for chronological issues and native speaker judgments
(More detailed data available upon request)

JUDGMENT BY NATIVE	PIVOT WORD	PART OF SPEECH	COLLOCATE	PART OF SPEECH
Inaccurately Flagged	low	j	vision	n
Inaccurately Flagged	school	n	counseling	n
Inaccurately Flagged	report	v	participant	n
Inaccurately Flagged	represent	v	text	n
Inaccurately Flagged	represent	v	equation	n
Inaccurately Flagged	line	n	equation	n
Inaccurately Flagged	limb	n	residual	j
Inaccurately Flagged	nerve	n	facial	j
Inaccurately Flagged	food	n	residuals	n
Inaccurately Flagged	set	v	result	n
Inaccurately Flagged	set	v	current	j
Inaccurately Flagged	current	j	search	n
Inaccurately Flagged	result	n	search	n
Inaccurately Flagged	sat	v	fat	n
Inaccurately Flagged	play	v	football	n
Inaccurately Flagged	winner	n	match	n
Inaccurately Flagged	video	n	clip	n
Inaccurately Flagged	begin	v	clip	n
Inaccurately Flagged	video	n	begin	v
Inaccurately Flagged	moment	n	break	n
Inaccurately Flagged	moment	n	commercial	j
Inaccurately Flagged	speak	v	interpreter	n
Inaccurately Flagged	continue	v	prime-time	n
Inaccurately Flagged	site	n	web	n
Inaccurately Flagged	phone	n	cell	n
Inaccurately Flagged	send	v	e-mail	n
Inaccurately Flagged	war	n	terror	n
Inaccurately Flagged	war	n	terrorism	n
Inaccurately Flagged	e-mail	n	address	n
Inaccurately Flagged	fiber	n	carbohydrate	n
Inaccurately Flagged	internet	n	use	v
Inaccurately Flagged	camera	n	digital	j
Inaccurately Flagged	internet	n	service	n
Inaccurately Flagged	internet	n	site	n
Inaccurately Flagged	search	n	engine	n
Inaccurately Flagged	get	v	e-mail	n
Inaccurately Flagged	show	n	reality	n
Inaccurately Flagged	visit	v	information	n
Inaccurately Flagged	phone	n	e-mail	n
Inaccurately Flagged	speed	n	mixer	n
Inaccurately Flagged	go	v	online	r
Inaccurately Flagged	room	n	chat	n
Inaccurately Flagged	check	v	site	n

Inaccurately Flagged	disorder	n	bipolar	j
Inaccurately Flagged	people	n	when	r
Inaccurately Flagged	independent	j	counsel	n
Inaccurately Flagged	serve	v	purpose	n
Inaccurately Flagged	best	j	player	n
Inaccurately Flagged	internet	n	company	n
Inaccurately Flagged	minute	n	preparation	n
Inaccurately Flagged	fat	n	sat	n
Inaccurately Flagged	seem	v	obvious	j
Inaccurately Flagged	attack	n	terror	n
Inaccurately Flagged	fight	v	terrorism	n
Inaccurately Flagged	e-mail	n	fax	v
Inaccurately Flagged	preparation	n	serving	n
Inaccurately Flagged	back	r	rock	v
Inaccurately Flagged	add	v	additional	j
Inaccurately Flagged	school	n	counselor	n
Inaccurately Flagged	visual	j	impairment	n
Inaccurately Flagged	result	n	current	j
Inaccurately Flagged	method	n	participant	n
Inaccurately Flagged	participant	n	complete	v
Inaccurately Flagged	student	n	impairment	n
Inaccurately Flagged	middle	j	ear	n
Inaccurately Flagged	indicate	v	participant	n
Inaccurately Flagged	symptom	n	depressive	n
Inaccurately Flagged	capital	n	social	j
Inaccurately Flagged	response	n	participant	n
Inaccurately Flagged	note	n	supra	n
Inaccurately Flagged	use	n	substance	n
Inaccurately Flagged	school	n	psychology	n
Inaccurately Flagged	equation	n	can	v
Inaccurately Flagged	note	n	text	n
Inaccurately Flagged	text	n	accompanying	j
Inaccurately Flagged	multiple	j	equation	n
Inaccurately Flagged	note	n	accompanying	j
Inaccurately Flagged	art	n	educator	n
Inaccurately Flagged	environmental	j	knowledge	n
Inaccurately Flagged	fat	n	protein	n
Inaccurately Flagged	cholesterol	n	carbohydrate	n
Inaccurately Flagged	lemon	n	zest	n
Inaccurately Flagged	total	j	minute	n
Inaccurately Flagged	salt	n	kosher	j
Inaccurately Flagged	beat	v	speed	n
Inaccurately Flagged	cook	v	spray	n
Inaccurately Flagged	article	n	copyright	n
Inaccurately Flagged	face	v	page	n
Inaccurately Flagged	fire	n	firefighter	n
Inaccurately Flagged	include	v	survivor	n
Inaccurately Flagged	scene	n	violence	n
Inaccurately Flagged	wife	n	survivor	n
Inaccurately Flagged	game	n	winner	n

Inaccurately Flagged	join	v	conversation	n
Inaccurately Flagged	join	v	phone	n
Inaccurately Flagged	report	n	tonight	r
Inaccurately Flagged	break	n	away	r
Inaccurately Flagged	commercial	j	away	r
Inaccurately Flagged	continue	v	commercial	j
Inaccurately Flagged	capital	n	gang	n
Inaccurately Flagged	attack	n	terrorist	j
Inaccurately Flagged	total	j	fat	n
Inaccurately Flagged	internet	n	access	n
Inaccurately Flagged	sheet	n	baking	n
Inaccurately Flagged	sugar	n	fiber	n
Inaccurately Flagged	message	n	e-mail	n
Inaccurately Flagged	type	n	diabetes	n
Inaccurately Flagged	phone	n	mobile	j
Inaccurately Flagged	can	v	e-mail	n
Inaccurately Flagged	information	n	site	n
Inaccurately Flagged	often	r	stir	v
Inaccurately Flagged	teaspoon	n	olive	j
Inaccurately Flagged	fiber	n	cholesterol	n
Inaccurately Flagged	food	n	organic	j
Inaccurately Flagged	message	n	text	n
Inaccurately Flagged	total	j	carbohydrate	n
Inaccurately Flagged	can	v	download	v
Inaccurately Flagged	can	v	online	r
Inaccurately Flagged	serving	n	minute	n
Inaccurately Flagged	will	v	end	v
Inaccurately Flagged	phone	n	cellular	j
Inaccurately Flagged	break	v	record	n
Inaccurately Flagged	system	n	expert	n
Inaccurately Flagged	welfare	n	recipient	n
Inaccurately Flagged	policy	n	industrial	j
Inaccurately Flagged	money	n	soft	j
Inaccurately Flagged	local	j	regional	j
Inaccurately Flagged	college	n	electoral	j
Inaccurately Flagged	low	j	lip	n
Inaccurately Flagged	scene	n	drug	n
Inaccurately Flagged	internet	n	connection	n
Inaccurately Flagged	insurance	n	deposit	n
Inaccurately Flagged	enter	v	week	n
Inaccurately Flagged	organize	v	help	v
Inaccurately Flagged	threat	n	terrorist	j
Inaccurately Flagged	art	n	education	n
Inaccurately Flagged	sexual	j	harassment	n
Inaccurately Flagged	weapon	n	destruction	n
Inaccurately Flagged	weapon	n	mass	j
Inaccurately Flagged	mass	j	destruction	n
Inaccurately Flagged	economic	j	reform	n
Inaccurately Flagged	program	n	nuclear	j
Inaccurately Flagged	reform	n	welfare	n

Inaccurately Flagged	high	j	heat	n
Inaccurately Flagged	price	n	gas	n
Inaccurately Flagged	campaign	n	finance	n
Inaccurately Flagged	world	n	wide	j
Inaccurately Flagged	waste	n	hazardous	j
Inaccurately Flagged	energy	n	renewable	j
Inaccurately Flagged	budget	n	balance	v
Inaccurately Flagged	social	j	network	n
Inaccurately Flagged	capital	n	gain	n
Inaccurately Flagged	personal	j	computer	n
Inaccurately Flagged	waste	n	solid	j
Inaccurately Flagged	site	n	visit	v
Inaccurately Flagged	financial	j	crisis	n
Inaccurately Flagged	defense	n	missile	n
Inaccurately Flagged	education	n	environmental	j
Inaccurately Flagged	school	n	charter	n
Inaccurately Flagged	end	n	cold	j
Inaccurately Flagged	campaign	n	reform	n
Inaccurately Flagged	image	n	body	n
Inaccurately Flagged	cut	v	spending	n
Inaccurately Flagged	loss	n	hearing	n
Inaccurately Flagged	vote	n	count	v
Inaccurately Flagged	acid	n	fatty	j
Inaccurately Flagged	sanction	n	economic	j
Inaccurately Flagged	special	j	prosecutor	n
Inaccurately Flagged	service	n	mental	j
Inaccurately Flagged	reform	n	finance	n
Inaccurately Flagged	policy	n	energy	n
Inaccurately Flagged	preparation	n	make	v
Inaccurately Flagged	use	v	cell	n
Inaccurately Flagged	vote	n	electoral	j
Inaccurately Flagged	share	v	story	n
Inaccurately Flagged	cheese	n	goat	n
Inaccurately Flagged	engineering	n	mechanical	j
Inaccurately Flagged	preparation	n	time	n
Inaccurately Flagged	trade	n	union	n
Inaccurately Flagged	management	n	waste	n
Inaccurately Flagged	whole	j	grain	n
Inaccurately Flagged	territory	n	occupied	j
Inaccurately Flagged	reduce	v	deficit	n
Inaccurately Flagged	burn	v	calorie	n
Inaccurately Flagged	defense	n	budget	n
Inaccurately Flagged	weapon	n	assault	n
Inaccurately Flagged	red	j	carpet	n
Inaccurately Flagged	official	n	intelligence	n
Inaccurately Flagged	sport	n	fan	n
Inaccurately Flagged	company	n	telephone	n
Inaccurately Flagged	care	n	manage	v
Inaccurately Flagged	reform	n	immigration	n
Inaccurately Flagged	organization	n	terrorist	j

Inaccurately Flagged	panel	n	solar	j
Inaccurately Flagged	office	n	accounting	n
Inaccurately Flagged	use	n	marijuana	n
Inaccurately Flagged	buy	v	best	r
Inaccurately Flagged	heavy	j	cream	n
Inaccurately Flagged	personal	j	trainer	n
Inaccurately Flagged	whole	j	food	n
Inaccurately Flagged	biological	j	chemical	n
Inaccurately Flagged	economic	j	downturn	n
Inaccurately Flagged	on	r	log	v
Inaccurately Flagged	page	n	home	n
Inaccurately Flagged	sea	n	salt	n
Inaccurately Flagged	waste	n	toxic	j
Inaccurately Flagged	continue	v	discussion	n
Inaccurately Flagged	green	j	tea	n
Inaccurately Flagged	market	n	drug	n
Inaccurately Flagged	time	n	appreciate	v
Inaccurately Flagged	people	n	welfare	n
Inaccurately Flagged	brain	n	injury	n
Inaccurately Flagged	justice	n	obstruction	n
Inaccurately Flagged	sport	n	utility	n
Inaccurately Flagged	political	j	democracy	n
Inaccurately Flagged	hearing	n	confirmation	n
Inaccurately Flagged	point	n	talking	n
Inaccurately Flagged	free	j	exercise	n
Inaccurately Flagged	consequence	n	unintended	j
Inaccurately Flagged	funny	j	really	r
Inaccurately Flagged	sugar	n	powdered	j
Inaccurately Flagged	program	n	welfare	n
Inaccurately Flagged	cool	j	really	r
Inaccurately Flagged	woman	n	battered	j
Inaccurately Flagged	intelligence	n	national	j
Inaccurately Flagged	statement	n	opening	n
Inaccurately Flagged	contribute	v	editor	n
Inaccurately Flagged	salt	n	sprinkle	v
Inaccurately Flagged	cut	n	spending	n
Inaccurately Flagged	system	n	cable	n
Inaccurately Flagged	prayer	n	school	n
Inaccurately Flagged	troop	n	number	n
Inaccurately Flagged	civilian	j	casualty	n
Inaccurately Flagged	national	j	gross	j
Inaccurately Flagged	special	j	class	n
Inaccurately Flagged	industry	n	computer	n
Inaccurately Flagged	wind	n	turbine	n
Inaccurately Flagged	hard	j	disk	n
Inaccurately Flagged	aid	n	hearing	n
Inaccurately Flagged	share	n	hold	v
Inaccurately Flagged	energy	n	wind	n
Inaccurately Flagged	height	n	inch	n
Inaccurately Flagged	girl	n	adolescent	j

Inaccurately Flagged	behind	r	child	n
Inaccurately Flagged	team	n	captain	n
Inaccurately Flagged	argument	n	closing	j
Inaccurately Flagged	road	n	block	n
Inaccurately Flagged	machine	n	fax	n
Inaccurately Flagged	line	n	telephone	n
Inaccurately Flagged	data	n	base	n
Inaccurately Flagged	war	n	oppose	v
Inaccurately Flagged	digital	j	use	v
Inaccurately Flagged	growth	n	revenue	n
Inaccurately Flagged	growth	n	income	n
Inaccurately Flagged	fat	n	diet	n
Inaccurately Flagged	gas	n	mask	n
Inaccurately Flagged	make	v	chocolate	n
Inaccurately Flagged	bone	n	density	n
Inaccurately Flagged	kill	v	attack	n
Inaccurately Flagged	nation	n	resolution	n
Inaccurately Flagged	major	j	championship	n
Inaccurately Flagged	camera	n	surveillance	n
Inaccurately Flagged	hard	j	currency	n
Inaccurately Flagged	tax	n	energy	n
Inaccurately Flagged	arm	n	sales	n
Inaccurately Flagged	board	n	message	n
Inaccurately Flagged	meeting	n	summit	n
Inaccurately Flagged	behavior	n	risky	j
Inaccurately Flagged	democratic	j	social	j
Inaccurately Flagged	election	n	official	n
Inaccurately Flagged	benefit	n	drug	n
Inaccurately Flagged	economic	j	aid	n
Inaccurately Flagged	sex	n	casual	j
Accurately Flagged	budget	n	amendment	n
Accurately Flagged	amendment	n	balanced	j
Accurately Flagged	suicide	n	bomber	n
Accurately Flagged	marriage	n	gay	j
Accurately Flagged	research	n	cell	n
Accurately Flagged	marriage	n	same-sex	j
Accurately Flagged	cell	n	embryonic	j
Accurately Flagged	stem	n	embryonic	j
Accurately Flagged	care	n	managed	j
Accurately Flagged	bill	n	crime	n
Accurately Flagged	package	n	stimulus	n
Accurately Flagged	research	n	stem	n
Accurately Flagged	suicide	n	bombing	n
Accurately Flagged	government	n	interim	j
Accurately Flagged	cell	n	stem	n
Accurately Flagged	fund	n	hedge	n
Accurately Flagged	health	n	reform	n
Accurately Flagged	budget	n	balanced	j
Accurately Flagged	care	n	reform	n
Accurately Flagged	party	n	reform	n

Accurately Flagged	force	n	coalition	n
Accurately Flagged	reduction	n	deficit	n
Accurately Flagged	new	j	millennium	n
Accurately Flagged	weapon	n	inspector	n
Accurately Flagged	bond	n	junk	n
Accurately Flagged	tax	n	flat	j
Accurately Flagged	lift	v	embargo	n
Accurately Flagged	reform	n	health-care	n
Accurately Flagged	ground	n	zero	n
Accurately Flagged	peace	n	conference	n
Accurately Flagged	force	n	allied	j
Accurately Flagged	agency	n	energy	n
Accurately Flagged	plan	n	peace	n
Accurately Flagged	central	j	command	n
Accurately Flagged	peace	n	accord	n
Accurately Flagged	arab	j	nation	n
Accurately Flagged	united	j	resolution	n
Accurately Flagged	saving	n	loan	n
Accurately Flagged	company	n	tobacco	n
Accurately Flagged	industry	n	tobacco	n
Accurately Flagged	rain	n	acid	n
Accurately Flagged	military	j	intervention	n
Accurately Flagged	biological	j	chemical	j
Accurately Flagged	change	n	regime	n
Accurately Flagged	war	n	ground	n
Accurately Flagged	special	j	operation	n
Accurately Flagged	missile	n	cruise	n
Accurately Flagged	layer	n	ozone	n
Accurately Flagged	security	n	airport	n
Accurately Flagged	coverage	n	universal	j
Accurately Flagged	bill	n	energy	n
Accurately Flagged	defense	n	air	n
Accurately Flagged	arm	n	embargo	n
Accurately Flagged	ballot	n	absentee	n
Accurately Flagged	vehicle	n	utility	n
Accurately Flagged	saving	n	bank	n
Accurately Flagged	disc	n	compact	j
Accurately Flagged	word	n	processor	n
Accurately Flagged	gain	n	tax	v
Accurately Flagged	troop	n	home	r
Accurately Flagged	cable	n	operator	n
Accurately Flagged	energy	n	crisis	n

Appendix 17. Five items not flagged by any parameters but deemed having chronological issues
(More detailed data available upon request)

PIVOT WORD	PART OF SPEECH	COLLOCATE	PART OF SPEECH	1990-94	1995-99	2000-04	2005-09
trade	n	deficit	n	424	354	121	174
federal	j	deficit	n	392	109	92	132
federal	j	insurance	n	345	84	66	139
land	n	reform	n	156	158	78	290
health	n	universal	j	169	62	95	279

Appendix 18. Items affected by colligational searches
(More detailed data available upon request)

MULTI-WORD UNIT

died cardinalnumber years
feet cardinalnumber inches
estimated cardinalnumber million people
estimated cardinalnumber percent
doubled in cardinalnumber years
down for cardinalnumber minutes
drive cardinalnumber miles
dropped cardinalnumber percent
earnings rose cardinalnumber percent
estimates that cardinalnumber percent
fall into cardinalnumber categories
fell cardinalnumber percent
fell cardinalnumber feet
founded cardinalnumber years
gained cardinalnumber pounds
gave up cardinalnumber runs
get cardinalnumber votes
given cardinalnumber months
go cardinalnumber miles
got cardinalnumber seconds
graduated cardinalnumber years
grow to cardinalnumber feet
have cardinalnumber minutes
hit cardinalnumber home runs
missed cardinalnumber games
hour cardinalnumber minutes
final cardinalnumber games
last cardinalnumber hours
founded cardinalnumber years
divided into cardinalnumber groups
about cardinalnumber meters
age cardinalnumber and older
approximately cardinalnumber minutes
least cardinalnumber feet
least cardinalnumber minutes
least cardinalnumber years
least cardinalnumber months
least cardinalnumber hours
maybe cardinalnumber minutes
nearly cardinalnumber decades
nearly cardinalnumber years
nearly cardinalnumber hours

nearly cardinalnumber weeks
past cardinalnumber months
increase in cardinalnumber years
increased cardinalnumber percent
jumped cardinalnumber percent
just cardinalnumber minutes
just cardinalnumber years ago
just cardinalnumber years
just cardinalnumber months
just cardinalnumber days
just cardinalnumber seconds
just cardinalnumber feet
just cardinalnumber miles
killed cardinalnumber people
last cardinalnumber years
lasted cardinalnumber days
launched cardinalnumber years
live cardinalnumber miles
located cardinalnumber miles
lost cardinalnumber pounds
makes cardinalnumber servings notes
married for cardinalnumber years
married cardinalnumber years
married with cardinalnumber children
maybe cardinalnumber hours
nearly cardinalnumber feet
nearly cardinalnumber percent
nearly cardinalnumber miles
nearly cardinalnumber months
only cardinalnumber feet
only cardinalnumber miles
only cardinalnumber minutes
only cardinalnumber percent
only cardinalnumber hours
only cardinalnumber inches
only cardinalnumber months
only cardinalnumber seconds
minutes cardinalnumber seconds
run cardinalnumber minutes
period of cardinalnumber years
period of cardinalnumber months
over cardinalnumber decades
over cardinalnumber years
past cardinalnumber days
out cardinalnumber minutes
over cardinalnumber hours
over cardinalnumber minutes

over cardinalnumber months
over cardinalnumber miles
over cardinalnumber feet
owns cardinalnumber percent
past cardinalnumber seasons
rate of cardinalnumber percent
read cardinalnumber pages
represent cardinalnumber percent
retired after cardinalnumber years
rose cardinalnumber percent
roughly cardinalnumber percent
roughly cardinalnumber years
run cardinalnumber miles
sentence of cardinalnumber years
served cardinalnumber years in prison
sit for cardinalnumber minutes
sold cardinalnumber million copies
covered cardinalnumber miles
dating back cardinalnumber years
about cardinalnumber inches
about cardinalnumber months
about cardinalnumber minutes
about cardinalnumber pounds
about cardinalnumber hours
about cardinalnumber years ago
about cardinalnumber seconds
about cardinalnumber feet
about cardinalnumber acres
about cardinalnumber cents
about cardinalnumber miles
about cardinalnumber percent
almost cardinalnumber years ago
almost cardinalnumber percent
almost cardinalnumber months
almost cardinalnumber feet
average of cardinalnumber percent
average of cardinalnumber years
approximately cardinalnumber years
approximately cardinalnumber percent
about cardinalnumber am
back in cardinalnumber minutes
below cardinalnumber percent
celebrating cardinalnumber years
city cardinalnumber miles
compared with cardinalnumber percent
declined cardinalnumber percent
died cardinalnumber years ago

died cardinalnumber months
almost cardinalnumber hours
almost cardinalnumber decades
almost cardinalnumber years
nearly cardinalnumber years ago
open cardinalnumber am
born cardinalnumber years
about cardinalnumber years
up to cardinalnumber miles
up to cardinalnumber percent
up to cardinalnumber hours
up to cardinalnumber minutes
about cardinalnumber yards
about cardinalnumber weeks
up to cardinalnumber months
there in cardinalnumber minutes
past cardinalnumber years
spend cardinalnumber minutes
spent cardinalnumber days
spent cardinalnumber years studying
spent cardinalnumber weeks
spent cardinalnumber years
survey cardinalnumber percent
take cardinalnumber minutes
took cardinalnumber years
took cardinalnumber hours
travel cardinalnumber miles
type cardinalnumber diabetes
wait cardinalnumber months
walked cardinalnumber miles
wall cardinalnumber feet
won cardinalnumber straight
about cardinalnumber square
beginning of the cardinalnumber century
driving meushimherthem crazy
gave meushimherthem a feeling
gave meushimherthem a tour
give meushimherthem a chance
give meushimherthem a hug
give meushimherthem a kiss
give meushimherthem a minute
give meushimherthem a second
give meushimherthem an hour
give meushimherthem strength
give meushimherthem an edge
give meushimherthem the tools
give meushimherthem a call

give meushimherthem a sense
give meushimherthem some insight
give meushimherthem a clue
give meushimherthem a little
give meushimherthem an advantage
help meushimherthem cope
make meushimherthem happy
remind meushimherthem how
tell meushimherthem how
give meushimherthem the benefit
gave meushimherthem a look
invited meushimherthem in
invited meushimherthem over
joining meushimherthem now
joining meushimherthem tonight
keep meushimherthem alive
keep meushimherthem busy
keep meushimherthem safe
keep meushimherthem healthy
keep meushimherthem informed
leave meushimherthem alone
let meushimherthem ask
let meushimherthem finish
let meushimherthem know
let meushimherthem just
made meushimherthem angry
made meushimherthem feel
made meushimherthem mad
made meushimherthem nervous
made meushimherthem think
made meushimherthem uncomfortable
made meushimherthem wonder
make meushimherthem proud
make meushimherthem safer
make meushimherthem more competitive
makes meushimherthem unique
make meushimherthem laugh
makes meushimherthem sad
help meushimherthem improve
gave meushimherthem a quick
hear meushimherthem sing
made meushimherthem sick
pulled meushimherthem aside
put meushimherthem in jail
see meushimherthem anymore
send meushimherthem a card
help meushimherthem out

give meushimherthem a break
gave meushimherthem confidence
gave meushimherthem the address
made meushimherthem cry
pushed meushimherthem away
help meushimherthem achieve
cheer meushimherthem up
help meushimherthem understand
make meushimherthem vulnerable
taught meushimherthem how
prove meushimherthem wrong
took meushimherthem aside
wish meushimherthem luck
early in the ordinalnumber century
finished ordinalnumber last year
early ordinalnumber century
end of the ordinalnumber century
making yourhishertheirmyourits ordinalnumber visit
marks the ordinalnumber anniversary
late ordinalnumber century
percent in the ordinalnumber quarter
came in ordinalnumber
grin on yourhishertheirmyourits face
gun in yourhishertheirmyourits hand
smile on yourhishertheirmyourits face
achieve yourhishertheirmyourits goals
affect yourhishertheirmyourits health
brushed yourhishertheirmyourits hair
cast yourhishertheirmyourits eyes
caught yourhishertheirmyourits eye
clapped yourhishertheirmyourits hands
closed yourhishertheirmyourits eyes
down yourhishertheirmyourits cheeks
down yourhishertheirmyourits throat
dragging yourhishertheirmyourits feet
dropped to yourhishertheirmyourits knees
early in yourhishertheirmyourits career
eat yourhishertheirmyourits vegetables
end yourhishertheirmyourits career
express yourhishertheirmyourits feelings
express yourhishertheirmyourits ideas
extended yourhishertheirmyourits hand
face in yourhishertheirmyourits hands
focus yourhishertheirmyourits efforts
follow yourhishertheirmyourits advice
follow yourhishertheirmyourits lead
forget yourhishertheirmyourits name

get yourhishertheirmyourits stuff
get on yourhishertheirmyourits nerves
grabbed yourhishertheirmyourits hand
grasped yourhishertheirmyourits hand
hands in yourhishertheirmyourits lap
hands in yourhishertheirmyourits pockets
hands over yourhishertheirmyourits ears
fell to yourhishertheirmyourits knees
got to yourhishertheirmyourits feet
hung yourhishertheirmyourits head
nodded yourhishertheirmyourits head
raised yourhishertheirmyourits glass
stuck yourhishertheirmyourits head
threw back yourhishertheirmyourits head
waved yourhishertheirmyourits hand
wiped yourhishertheirmyourits hands
hear yourhishertheirmyourits voice
heard yourhishertheirmyourits voice
held yourhishertheirmyourits breath
hid yourhishertheirmyourits face
hit yourhishertheirmyourits head
hurt yourhishertheirmyourits feelings
image in yourhishertheirmyourits mind
broke yourhishertheirmyourits heart
jumped to yourhishertheirmyourits feet
keep yourhishertheirmyourits distance
kissed yourhishertheirmyourits hand
lifted yourhishertheirmyourits foot
opened yourhishertheirmyourits eyes
palm of yourhishertheirmyourits hand
pay yourhishertheirmyourits employees
poked yourhishertheirmyourits head
popped into yourhishertheirmyourits head
pursue yourhishertheirmyourits goals
gun to yourhishertheirmyourits head
raised yourhishertheirmyourits hand
rested yourhishertheirmyourits head
rose to yourhishertheirmyourits feet
scratching yourhishertheirmyourits heads
shifted yourhishertheirmyourits focus
shut yourhishertheirmyourits eyes
slapped yourhishertheirmyourits hand
snapped yourhishertheirmyourits fingers
tears streaming down yourhishertheirmyourits face
corner of yourhishertheirmyourits eye
expression on yourhishertheirmyourits face
tips of yourhishertheirmyourits fingers

top of yourhishertheirmyourits head
finish yourhishertheirmyourits work
get yourhishertheirmyourits attention
give yourhishertheirmyourits opinion
wash yourhishertheirmyourits hands
wash yourhishertheirmyourits face
whispered in yourhishertheirmyourits ear
wiped yourhishertheirmyourits eyes
wiped yourhishertheirmyourits face
tears in yourhishertheirmyourits eyes
called yourhishertheirmyourits name
devoted yourhishertheirmyourits life
elbows on yourhishertheirmyourits knees
down on yourhishertheirmyourits knees
back of yourhishertheirmyourits mind
keep yourhishertheirmyourits promise
keep yourhishertheirmyourits feet
keep yourhishertheirmyourits mouth shut
kill yourhishertheirmyourits wife
killed yourhishertheirmyourits husband
killed yourhishertheirmyourits brother
kissed yourhishertheirmyourits cheek
knife in yourhishertheirmyourits hand
lay yourhishertheirmyourits eggs
learn from yourhishertheirmyourits mistakes
left yourhishertheirmyourits mark
listen to yourhishertheirmyourits voice
live yourhishertheirmyourits life
looked at yourhishertheirmyourits watch
looking in yourhishertheirmyourits direction
lost yourhishertheirmyourits balance
lost yourhishertheirmyourits job
lying on yourhishertheirmyourits side
made yourhishertheirmyourits debut
made yourhishertheirmyourits way
made up yourhishertheirmyourits mind
make yourhishertheirmyourits pitch
mention yourhishertheirmyourits name
money in yourhishertheirmyourits pocket
doubt in yourhishertheirmyourits mind
opened yourhishertheirmyourits mouth
off yourhishertheirmyourits debt
percent of yourhishertheirmyourits income
protect yourhishertheirmyourits interests
pursue yourhishertheirmyourits interests
quit yourhishertheirmyourits job
risking yourhishertheirmyourits life

ruin yourhishertheirmyourits life
saved yourhishertheirmyourits life
stretch yourhishertheirmyourits legs
off yourhishertheirmyourits shoes
off yourhishertheirmyourits clothes
off yourhishertheirmyourits hat
off yourhishertheirmyourits shirt
look in yourhishertheirmyourits eyes
rest of yourhishertheirmyourits life
made yourhishertheirmyourits mark
meet yourhishertheirmyourits goal
promote yourhishertheirmyourits new
workers lost yourhishertheirmyourits jobs
wrote in yourhishertheirmyourits journal
rolled yourhishertheirmyourits head
over yourhishertheirmyourits shoulder
pay off yourhishertheirmyourits credit
pay yourhishertheirmyourits bills
placed yourhishertheirmyourits hand
pulling yourhishertheirmyourits leg
put yourhishertheirmyourits hand
put yourhishertheirmyourits arm
put yourhishertheirmyourits clothes
put yourhishertheirmyourits finger
put yourhishertheirmyourits shoes
raised yourhishertheirmyourits arms
raised yourhishertheirmyourits head
raised yourhishertheirmyourits voice
reach yourhishertheirmyourits goal
reached into yourhishertheirmyourits pocket
reached yourhishertheirmyourits peak
read yourhishertheirmyourits mind
reveal yourhishertheirmyourits secrets
roll up yourhishertheirmyourits sleeves
said yourhishertheirmyourits voice
see yourhishertheirmyourits face
set yourhishertheirmyourits sights
share yourhishertheirmyourits concerns
sign yourhishertheirmyourits name
sit on yourhishertheirmyourits lap
sitting at yourhishertheirmyourits desk
pay yourhishertheirmyourits debts
pay yourhishertheirmyourits respects
perform yourhishertheirmyourits duties
protect yourhishertheirmyourits privacy
prove yourhishertheirmyourits point
back on yourhishertheirmyourits feet

mind yourhishertheirmyourits own business
check out yourhishertheirmyourits web site
check yourhishertheirmyourits local
catch yourhishertheirmyourits breath
caught yourhishertheirmyourits attention
change yourhishertheirmyourits mind
cleared yourhishertheirmyourits throat
threw yourhishertheirmyourits arms
affect yourhishertheirmyourits ability
argue yourhishertheirmyourits case
arms at yourhishertheirmyourits sides
arms over yourhishertheirmyourits head
around yourhishertheirmyourits neck
based on yourhishertheirmyourits experience
began yourhishertheirmyourits career
blew yourhishertheirmyourits nose
blood on yourhishertheirmyourits face
broke yourhishertheirmyourits leg
change yourhishertheirmyourits behavior
change yourhishertheirmyourits attitude
changed yourhishertheirmyourits name
changed yourhishertheirmyourits position
changed yourhishertheirmyourits life
consider yourhishertheirmyourits options
continue yourhishertheirmyourits conversation
continue yourhishertheirmyourits discussion
covered yourhishertheirmyourits mouth
covered yourhishertheirmyourits face
cut yourhishertheirmyourits hair
describe yourhishertheirmyourits experience
destroyed yourhishertheirmyourits life
direct yourhishertheirmyourits attention
educate yourhishertheirmyourits children
focus yourhishertheirmyourits attention
grabbed yourhishertheirmyourits arm
back of yourhishertheirmyourits neck
yourhishertheirmyourits most recent book
the trunk of yourhishertheirmyourits car
increase yourhishertheirmyourits chances
influenced yourhishertheirmyourits decision
keep yourhishertheirmyourits balance
keep yourhishertheirmyourits cool
leaned back in yourhishertheirmyourits chair
marry yourhishertheirmyourits daughter
pack yourhishertheirmyourits bags
percent of yourhishertheirmyourits budget
sipped yourhishertheirmyourits coffee

consequences of yourhishertheirmyourits actions
achieve yourhishertheirmyourits objectives
use yourhishertheirmyourits credit card
waving yourhishertheirmyourits arms
wrapped yourhishertheirmyourits arms
swinging yourhishertheirmyourits arms
up yourhishertheirmyourits sleeves
turned yourhishertheirmyourits back
taken yourhishertheirmyourits toll
took yourhishertheirmyourits arm
wore yourhishertheirmyourits hair
stopped in yourhishertheirmyourits tracks
take yourhishertheirmyourits medicine
tears streaming down yourhishertheirmyourits cheeks
taking yourhishertheirmyourits call
threw yourhishertheirmyourits arms around
threw up yourhishertheirmyourits hands
took yourhishertheirmyourits leave
took yourhishertheirmyourits hand
walking yourhishertheirmyourits dog
want yourhishertheirmyourits kids
keep yourhishertheirmyourits back straight
share yourhishertheirmyourits views
sound of yourhishertheirmyourits voice
spent yourhishertheirmyourits career
spent yourhishertheirmyourits entire
stuck out yourhishertheirmyourits hand
support yourhishertheirmyourits position
take yourhishertheirmyourits advice
thing on yourhishertheirmyourits mind
took off yourhishertheirmyourits coat
took off yourhishertheirmyourits hat
touched yourhishertheirmyourits face
turned yourhishertheirmyourits head
turned yourhishertheirmyourits attention
up from yourhishertheirmyourits chair
up from yourhishertheirmyourits desk
visit yourhishertheirmyourits family
wrote in yourhishertheirmyourits memoir
pregnant with yourhishertheirmyourits ordinalnumber child
released yourhishertheirmyourits ordinalnumber album
put iyouheshetheyitwe in yourhishertheirmyourits pocket
gave meushimherthem yourhishertheirmyourits card
celebrated yourhishertheirmyourits ordinalnumber birthday
celebrates yourhishertheirmyourits ordinalnumber anniversary
birth of yourhishertheirmyourits ordinalnumber child
poured myselfyourselfhimselfherselfitselfyourselvesthemselves a glass

handed iyouheshetheyitwe back
do iyouheshetheyitwe agree
everywhere iyouheshetheyitwe go
frankly iyouheshetheyitwe don
frankly iyouheshetheyitwe think
give iyouheshetheyitwe a hint
give iyouheshetheyitwe an example
got what iyouheshetheyitwe deserved
help iyouheshetheyitwe find
hopefully iyouheshetheyitwe will
hours iyouheshetheyitwe had
how iyouheshetheyitwe feel
how iyouheshetheyitwe differs
how iyouheshetheyitwe evolved
how iyouheshetheyitwe fit
how iyouheshetheyitwe happened
how iyouheshetheyitwe works
how much iyouheshetheyitwe hated
how much iyouheshetheyitwe costs
how iyouheshetheyitwe interact
how iyouheshetheyitwe might
how iyouheshetheyitwe perceive
how iyouheshetheyitwe relate
how iyouheshetheyitwe view
how iyouheshetheyitwe define
find iyouheshetheyitwe hard
realized iyouheshetheyitwe had
think iyouheshetheyitwe s fair
wish iyouheshetheyitwe had
instead iyouheshetheyitwe found
make iyouheshetheyitwe difficult
make iyouheshetheyitwe easier
months iyouheshetheyitwe had
food iyouheshetheyitwe eat
give iyouheshetheyitwe a try
give iyouheshetheyitwe an idea
tomorrow iyouheshetheyitwe re going
well iyouheshetheyitwe guess
when iyouheshetheyitwe died
when iyouheshetheyitwe got
where iyouheshetheyitwe hid
where iyouheshetheyitwe got
wherever iyouheshetheyitwe go
anymore iyouheshetheyitwe just
before iyouheshetheyitwe met
feared iyouheshetheyitwe might
guy iyouheshetheyitwe know

keep iyouheshetheyitwe clean
like iyouheshetheyitwe know
like iyouheshetheyitwe know
made iyouheshetheyitwe clear
made iyouheshetheyitwe impossible
make iyouheshetheyitwe a priority
make iyouheshetheyitwe fast
make iyouheshetheyitwe harder
make iyouheshetheyitwe illegal
make iyouheshetheyitwe interesting
make iyouheshetheyitwe more
make iyouheshetheyitwe more efficient
makes iyouheshetheyitwe hard
man iyouheshetheyitwe love
exactly iyouheshetheyitwe mean
months after iyouheshetheyitwe met
now iyouheshetheyitwe realize
smile iyouheshetheyitwe looked
techniques iyouheshetheyitwe learned
skills iyouheshetheyitwe learned
then iyouheshetheyitwe kissed
knock iyouheshetheyitwe off
make iyouheshetheyitwe safe
way iyouheshetheyitwe look
where iyouheshetheyitwe live
position iyouheshetheyitwe held
argue that iyouheshetheyitwe should
first iyouheshetheyitwe seemed
sorry iyouheshetheyitwe said
later iyouheshetheyitwe returned
maybe iyouheshetheyitwe should
minute iyouheshetheyitwe said
mirror iyouheshetheyitwe saw
months later iyouheshetheyitwe received
okay iyouheshetheyitwe said
see iyouheshetheyitwe again
see iyouheshetheyitwe tomorrow
take iyouheshetheyitwe personally
when iyouheshetheyitwe saw
when iyouheshetheyitwe returned
whenever iyouheshetheyitwe see
iyouheshetheyitwe can t bear
couldn t take iyouheshetheyitwe anymore
guess iyouheshetheyitwe could
absolutely iyouheshetheyitwe mean
fast as iyouheshetheyitwe could
fast as iyouheshetheyitwe can

quickly as iyouheshetheyitwe could
buy iyouheshetheyitwe a drink
confident iyouheshetheyitwe can
day iyouheshetheyitwe arrived
hope iyouheshetheyitwe can
suppose iyouheshetheyitwe could
wish iyouheshetheyitwe could
maybe iyouheshetheyitwe could
otherwise iyouheshetheyitwe could
realized iyouheshetheyitwe could
sure iyouheshetheyitwe can
air iyouheshetheyitwe breathe
attention iyouheshetheyitwe deserves
best iyouheshetheyitwe could
best iyouheshetheyitwe can
day iyouheshetheyitwe died
day iyouheshetheyitwe was born
very much iyouheshetheyitwe appreciate
afraid iyouheshetheyitwe might
appreciate iyouheshetheyitwe coming
when iyouheshetheyitwe arrived
when iyouheshetheyitwe came
whenever iyouheshetheyitwe came
whenever iyouheshetheyitwe can
where iyouheshetheyitwe came
where iyouheshetheyitwe belong
why iyouheshetheyitwe chose
wondering if iyouheshetheyitwe could
take iyouheshetheyitwe easy
take iyouheshetheyitwe in stride
afraid iyouheshetheyitwe ll
well iyouheshetheyitwe think
anyway iyouheshetheyitwe want
anyway iyouheshetheyitwe ll
take iyouheshetheyitwe anymore
think iyouheshetheyitwe s kind
minute iyouheshetheyitwe start
then iyouheshetheyitwe hit
minutes iyouheshetheyitwe would
mistake iyouheshetheyitwe think
announced iyouheshetheyitwe would
else iyouheshetheyitwe want
anywhere iyouheshetheyitwe want
assured meushimherthem that iyouheshetheyitwe would
dollars iyouheshetheyitwe took
dress iyouheshetheyitwe wore
exactly what iyouheshetheyitwe wanted

fear iyouheshetheyitwe will
feared iyouheshetheyitwe would
figured iyouheshetheyitwe would
promised iyouheshetheyitwe would
how iyouheshetheyitwe treat
believe iyouheshetheyitwe will
decided iyouheshetheyitwe wanted
think iyouheshetheyitwe s appropriate
doubt iyouheshetheyitwe will
guess iyouheshetheyitwe ll
promise iyouheshetheyitwe won t
suppose iyouheshetheyitwe would
think iyouheshetheyitwe s smart
think iyouheshetheyitwe probably
think iyouheshetheyitwe will
think iyouheshetheyitwe s important
think iyouheshetheyitwe s interesting
think iyouheshetheyitwe s wonderful
think iyouheshetheyitwe s pretty
think iyouheshetheyitwe deserve
think iyouheshetheyitwe ought
thought iyouheshetheyitwe might
thought iyouheshetheyitwe would
thought iyouheshetheyitwe was funny
wish iyouheshetheyitwe would
ll tell iyouheshetheyitwe why
sure iyouheshetheyitwe ll
later iyouheshetheyitwe will
maybe iyouheshetheyitwe ll
otherwise iyouheshetheyitwe will
perhaps iyouheshetheyitwe will
personally iyouheshetheyitwe think
said iyouheshetheyitwe will
afraid iyouheshetheyitwe would
so iyouheshetheyitwe guess
sometimes iyouheshetheyitwe wonder
subjects iyouheshetheyitwe teach
sure iyouheshetheyitwe understand
take iyouheshetheyitwe back
thank iyouheshetheyitwe all
thank iyouheshetheyitwe for joining
thank iyouheshetheyitwe sir
thank iyouheshetheyitwe so
the way iyouheshetheyitwe dress
then iyouheshetheyitwe smiled
then iyouheshetheyitwe started
then iyouheshetheyitwe noticed

then iyouheshetheyitwe realized
thing iyouheshetheyitwe ve ever
stupid iyouheshetheyitwe think
wrong iyouheshetheyitwe think
tomorrow iyouheshetheyitwe ll
minutes iyouheshetheyitwe will
way iyouheshetheyitwe operate
way iyouheshetheyitwe s supposed
well iyouheshetheyitwe certainly
when iyouheshetheyitwe started
when iyouheshetheyitwe woke
whenever iyouheshetheyitwe want
where iyouheshetheyitwe stood
why do iyouheshetheyitwe think
why would iyouheshetheyitwe want
wondered if iyouheshetheyitwe would
worried that iyouheshetheyitwe might
thing iyouheshetheyitwe heard
give iyouheshetheyitwe cardinalnumber dollars

Appendix 19. Sample of 100 MWUs highlighting the percentage which were extended beyond their strings
(More detailed data available upon request)

WAS MWU EXTENDED OR NOT	MOST FREQUENT MWU	FINAL MWU AFTER CONSIDERING WHETHER TO EXTEND IT OR NOT
yes	balancing act	a balancing act
yes	class action	a class action lawsuit
yes	more active	a more active role in
yes	very active	a very active
yes	about cardinalnumber inches	about 10 inches long
yes	about cardinalnumber years ago	about 3 years ago I
yes	about cardinalnumber acres	about 50 acres of
yes	about cardinalnumber percent	about 50 percent of the
yes	absolutely iyouheshetheyitwe mean	absolutely I mean I
yes	accept responsibility	accept responsibility for
yes	accepted an invitation	accepted an invitation to
yes	account for about	account for about 10 percent of
yes	account for only	account for only 20 percent
yes	affect yourhishertheirmyourits ability	affect their ability to
yes	active member	an active member of the
yes	willing to accept	are willing to accept the
yes	able to afford	be able to afford
yes	able to handle	be able to handle the
yes	abuse and neglect	child abuse and neglect
yes	denied access	denied access to
yes	equal access	equal access to
yes	act together	get their act together
yes	given access	given access to
yes	quickly added	he added quickly
yes	refused to accept	I refused to accept the
yes	never be able to	I'll never be able to
yes	widely accepted	is widely accepted as
yes	lasted about	lasted about 20
yes	limited ability	limited ability to
yes	more acceptable	more acceptable to
yes	active duty	on active duty in
yes	abuse cases	sexual abuse cases
yes	abuse problems	substance abuse problems
yes	legal action	take legal action against
yes	academic community	the academic community
yes	academic year	the academic year
yes	consequences of yourhishertheirmyourits actions	the consequences of their actions
yes	man accused	the man accused of
yes	most active	the most active

yes	need for additional	the need for additional
yes	daily activities	their daily activities
yes	accept the fact	to accept the fact that
yes	achieve yourhishertheirmyourits objectives	to achieve their objectives
yes	bring about	to bring about
yes	gain access	to gain access to
yes	get across	to get across
yes	help meushimherthem achieve	to help them achieve
yes	take action	to take action
yes	worry about	to worry about
yes	must accept	we must accept the
yes	willing and able	willing and able to
yes	may be able	you may be able to
yes	might be able	you might be able to
no	abortion clinics	abortion clinics
no	about cardinalnumber months	about 2 months
no	about cardinalnumber minutes	about 20 minutes
no	about cardinalnumber pounds	about 20 pounds
no	about cardinalnumber hours	about 3 hours
no	about cardinalnumber seconds	about 30 seconds
no	about cardinalnumber feet	about 5 feet
no	about cardinalnumber cents	about 50 cents
no	about cardinalnumber miles	about 50 miles
no	access to education	access to education
no	accomplish things	accomplish things
no	acts committed	acts committed
no	actually pretty	actually pretty
no	ad campaign	ad campaign
no	adapt to new	adapt to new
no	add extra	add extra
no	added to the list	added to the list
no	adds another dimension	adds another dimension
no	alcohol abuse	alcohol abuse
no	become active	become active
no	classroom activities	classroom activities
no	company acquired	company acquired
no	cost about	cost about
no	course of action	course of action
no	cards accepted	credit cards accepted
no	criminal activity	criminal activity
no	engage in activities	engage in activities
no	extracurricular activities	extracurricular activities
no	high achievement	high achievement
no	involved in activities	involved in activities
no	just about	just about
no	learning activities	learning activities
no	newspaper ads	newspaper ads
no	participate in activities	participate in activities
no	physical activity	physical activity
no	political action	political action

no	political activists	political activists
no	reservations and credit cards accepted	reservations and credit cards accepted
no	retirement accounts	retirement accounts
no	running ads	running ads
no	sexual activity	sexual activity
no	sexually abused	sexually abused
no	sexually active	sexually active
no	skills and abilities	skills and abilities
no	sports activities	sports activities
no	television ads	television ads
no	thanks for coming	thanks for coming

Appendix 20. Inter-rater differences for semantic transparency ratings

MWU	RATING	REVIEWER
a dream come true	**6**	Reviewer 1
a dream come true	12	Reviewer 2
a foster home	**6**	Reviewer 1
a foster home	8	Reviewer 2
a goldmine of	**4**	Reviewer 1
a goldmine of	0	Reviewer 2
a good night's sleep	**6**	Reviewer 1
a good night's sleep	12	Reviewer 2
a great distance	**6**	Reviewer 1
a great distance	**4**	Reviewer 2
a line of credit	12	Reviewer 1
a line of credit	8	Reviewer 2
a machine gun	**6**	Reviewer 1
a machine gun	8	Reviewer 2
a means to an end	**4**	Reviewer 1
a means to an end	**6**	Reviewer 2
a mobile home	**6**	Reviewer 1
a mobile home	8	Reviewer 2
a pair of jeans	**6**	Reviewer 1
a pair of jeans	8	Reviewer 2
a piece of furniture	**6**	Reviewer 1
a piece of furniture	8	Reviewer 2
a piece of legislation	**6**	Reviewer 1
a piece of legislation	8	Reviewer 2
a piece of music	**6**	Reviewer 1
a piece of music	8	Reviewer 2
a plastic surgeon	**6**	Reviewer 1
a plastic surgeon	8	Reviewer 2
a step further	**4**	Reviewer 1
a step further	0	Reviewer 2
acting out	8	Reviewer 1
acting out	**4**	Reviewer 2
all-expenses paid	**6**	Reviewer 1
all-expenses paid	8	Reviewer 2
an hour's drive	**6**	Reviewer 1
an hour's drive	8	Reviewer 2
and possibly even	12	Reviewer 1
and possibly even	8	Reviewer 2
and so forth	**4**	Reviewer 1
and so forth	0	Reviewer 2
at the end of the day	**4**	Reviewer 1
at the end of the day	0	Reviewer 2
begs the question	8	Reviewer 1
begs the question	**6**	Reviewer 2
brush off	**4**	Reviewer 1
brush off	0	Reviewer 2

cases filed	**6**	Reviewer 1
cases filed	8	Reviewer 2
chain reaction	8	Reviewer 1
chain reaction	**6**	Reviewer 2
change of direction	12	Reviewer 1
change of direction	8	Reviewer 2
children adopted from	**6**	Reviewer 1
children adopted from	8	Reviewer 2
come full circle	**4**	Reviewer 1
come full circle	0	Reviewer 2
commander in chief	**6**	Reviewer 1
commander in chief	**4**	Reviewer 2
community college	**6**	Reviewer 1
community college	8	Reviewer 2
consumer reports	**6**	Reviewer 1
consumer reports	12	Reviewer 2
continuing education	**6**	Reviewer 1
continuing education	8	Reviewer 2
cranked up	8	Reviewer 1
cranked up	**6**	Reviewer 2
credit report	**6**	Reviewer 1
credit report	12	Reviewer 2
dollar bill	**6**	Reviewer 1
dollar bill	8	Reviewer 2
dragging their feet	**4**	Reviewer 1
dragging their feet	0	Reviewer 2
drift off	**4**	Reviewer 1
drift off	0	Reviewer 2
driving me crazy	8	Reviewer 1
driving me crazy	**6**	Reviewer 2
due process	8	Reviewer 1
due process	**6**	Reviewer 2
earned a master's degree in	**6**	Reviewer 1
earned a master's degree in	8	Reviewer 2
employee benefits	**6**	Reviewer 1
employee benefits	8	Reviewer 2
every so often	12	Reviewer 1
every so often	**6**	Reviewer 2
falling behind	8	Reviewer 1
falling behind	**4**	Reviewer 2
family planning	**4**	Reviewer 1
family planning	12	Reviewer 2
figure out a way to	8	Reviewer 1
figure out a way to	**6**	Reviewer 2
foot in the door	**4**	Reviewer 1
foot in the door	0	Reviewer 2
from top to bottom	12	Reviewer 1
from top to bottom	**4**	Reviewer 2
gas station	12	Reviewer 1
gas station	8	Reviewer 2

gave up four runs	8	Reviewer 1
gave up four runs	**6**	Reviewer 2
get a laugh	8	Reviewer 1
get a laugh	**6**	Reviewer 2
get caught up in	**4**	Reviewer 1
get caught up in	0	Reviewer 2
get hooked on	8	Reviewer 1
get hooked on	**6**	Reviewer 2
get kicked out of	8	Reviewer 1
get kicked out of	**6**	Reviewer 2
get the hell out of here	8	Reviewer 1
get the hell out of here	**6**	Reviewer 2
give me a minute	8	Reviewer 1
give me a minute	**6**	Reviewer 2
give me a second	8	Reviewer 1
give me a second	**6**	Reviewer 2
give me an hour	8	Reviewer 1
give me an hour	**6**	Reviewer 2
give rise to the	**4**	Reviewer 1
give rise to the	0	Reviewer 2
given up hope	8	Reviewer 1
given up hope	**6**	Reviewer 2
go to great lengths to	8	Reviewer 1
go to great lengths to	**6**	Reviewer 2
going forward	12	Reviewer 1
going forward	**6**	Reviewer 2
good taste	8	Reviewer 1
good taste	12	Reviewer 2
graduate programs	8	Reviewer 1
graduate programs	12	Reviewer 2
guys get	12	Reviewer 1
guys get	8	Reviewer 2
had somehow	12	Reviewer 1
had somehow	8	Reviewer 2
hang around	8	Reviewer 1
hang around	12	Reviewer 2
hard rock	**6**	Reviewer 1
hard rock	**4**	Reviewer 2
has miles of	12	Reviewer 1
has miles of	**6**	Reviewer 2
hate crimes	8	Reviewer 1
hate crimes	12	Reviewer 2
he didn't even bother to	12	Reviewer 1
he didn't even bother to	**6**	Reviewer 2
he got nowhere	8	Reviewer 1
he got nowhere	**6**	Reviewer 2
he got to his feet	8	Reviewer 1
he got to his feet	**6**	Reviewer 2
he hung his head	8	Reviewer 1
he hung his head	**6**	Reviewer 2

he threw back his head and	8	Reviewer 1
he threw back his head and	**6**	Reviewer 2
head back	8	Reviewer 1
head back	**6**	Reviewer 2
head down to	8	Reviewer 1
head down to	**6**	Reviewer 2
head out	12	Reviewer 1
head out	8	Reviewer 2
heart attacks and strokes	**6**	Reviewer 1
heart attacks and strokes	12	Reviewer 2
her eyes lit up	8	Reviewer 1
her eyes lit up	**6**	Reviewer 2
her face lit up	8	Reviewer 1
her face lit up	**6**	Reviewer 2
his eyes darted	8	Reviewer 1
his eyes darted	**6**	Reviewer 2
his heart racing	8	Reviewer 1
his heart racing	**6**	Reviewer 2
his index finger	**6**	Reviewer 1
his index finger	8	Reviewer 2
his little finger	**6**	Reviewer 1
his little finger	8	Reviewer 2
hold elections	**6**	Reviewer 1
hold elections	8	Reviewer 2
home care	**6**	Reviewer 1
home care	12	Reviewer 2
how wonderful it is	12	Reviewer 1
how wonderful it is	**6**	Reviewer 2
however remains	12	Reviewer 1
however remains	**6**	Reviewer 2
I fell in love with	8	Reviewer 1
I fell in love with	**6**	Reviewer 2
I felt somehow	12	Reviewer 1
I felt somehow	**6**	Reviewer 2
I find it hard to	8	Reviewer 1
I find it hard to	**6**	Reviewer 2
I get bored	**6**	Reviewer 1
I get bored	8	Reviewer 2
I get home	**6**	Reviewer 1
I get home	8	Reviewer 2
I get the impression that	**6**	Reviewer 1
I get the impression that	8	Reviewer 2
I got here	**6**	Reviewer 1
I got here	8	Reviewer 2
I hardly ever	12	Reviewer 1
I hardly ever	8	Reviewer 2
I have a feeling	8	Reviewer 1
I have a feeling	**6**	Reviewer 2
I wish I had	12	Reviewer 1
I wish I had	8	Reviewer 2

I wonder how	12	Reviewer 1
I wonder how	**6**	Reviewer 2
I would consider	12	Reviewer 1
I would consider	0	Reviewer 2
illegal aliens	**6**	Reviewer 1
illegal aliens	12	Reviewer 2
imagine how it	12	Reviewer 1
imagine how it	**6**	Reviewer 2
in punitive damages	8	Reviewer 1
in punitive damages	12	Reviewer 2
in the back yard	12	Reviewer 1
in the back yard	12	Reviewer 2
in the long run	8	Reviewer 1
in the long run	**4**	Reviewer 2
in vain	8	Reviewer 1
in vain	0	Reviewer 2
intellectual property	**6**	Reviewer 1
intellectual property	12	Reviewer 2
interestingly enough	8	Reviewer 1
interestingly enough	**6**	Reviewer 2
is gaining momentum	12	Reviewer 1
is gaining momentum	**6**	Reviewer 2
it will ever	12	Reviewer 1
it will ever	8	Reviewer 2
just a little bit	12	Reviewer 1
just a little bit	8	Reviewer 2
kept at bay	8	Reviewer 1
kept at bay	**4**	Reviewer 2
kept in check	8	Reviewer 1
kept in check	**4**	Reviewer 2
late and early	12	Reviewer 1
late and early	0	Reviewer 2
liberal arts college	8	Reviewer 1
liberal arts college	12	Reviewer 2
long stretches	8	Reviewer 1
long stretches	**4**	Reviewer 2
love how	12	Reviewer 1
love how	**6**	Reviewer 2
major political parties	**6**	Reviewer 1
major political parties	8	Reviewer 2
make yourself comfortable	8	Reviewer 1
make yourself comfortable	**4**	Reviewer 2
moral authority	**6**	Reviewer 1
moral authority	8	Reviewer 2
more than ever	12	Reviewer 1
more than ever	**4**	Reviewer 2
most definitely	**6**	Reviewer 1
most definitely	**4**	Reviewer 2
never hurt	12	Reviewer 1
never hurt	**6**	Reviewer 2

new hires	12	Reviewer 1
new hires	8	Reviewer 2
no matter how much	12	Reviewer 1
no matter how much	**6**	Reviewer 2
nodded in agreement	8	Reviewer 1
nodded in agreement	12	Reviewer 2
not even close	12	Reviewer 1
not even close	**6**	Reviewer 2
nuclear arms	**6**	Reviewer 1
nuclear arms	**4**	Reviewer 2
nuclear program	**6**	Reviewer 1
nuclear program	12	Reviewer 2
oddly enough	8	Reviewer 1
oddly enough	**6**	Reviewer 2
off in the direction of	8	Reviewer 1
off in the direction of	**6**	Reviewer 2
off in the distance	8	Reviewer 1
off in the distance	**6**	Reviewer 2
off to a good start	8	Reviewer 1
off to a good start	**6**	Reviewer 2
oral history	**6**	Reviewer 1
oral history	12	Reviewer 2
organized crime	**6**	Reviewer 1
organized crime	8	Reviewer 2
party leaders	**6**	Reviewer 1
party leaders	8	Reviewer 2
party members	**6**	Reviewer 1
party members	8	Reviewer 2
pay benefits	**6**	Reviewer 1
pay benefits	8	Reviewer 2
percent of gross domestic product	**6**	Reviewer 1
percent of gross domestic product	12	Reviewer 2
piece of information that	**6**	Reviewer 1
piece of information that	8	Reviewer 2
plastic surgery	**6**	Reviewer 1
plastic surgery	8	Reviewer 2
poked his head	12	Reviewer 1
poked his head	8	Reviewer 2
political parties	**6**	Reviewer 1
political parties	8	Reviewer 2
popped into my head	8	Reviewer 1
popped into my head	**6**	Reviewer 2
pretty soon	12	Reviewer 1
pretty soon	8	Reviewer 2
pretty well	12	Reviewer 1
pretty well	8	Reviewer 2
profit margins	**6**	Reviewer 1
profit margins	12	Reviewer 2
public housing	**6**	Reviewer 1
public housing	12	Reviewer 2

public servant	8	Reviewer 1
public servant	12	Reviewer 2
pulling the strings	**4**	Reviewer 1
pulling the strings	0	Reviewer 2
put to rest	**4**	Reviewer 1
put to rest	0	Reviewer 2
quality control	**6**	Reviewer 1
quality control	8	Reviewer 2
rained down	12	Reviewer 1
rained down	**6**	Reviewer 2
raised an eyebrow	**4**	Reviewer 1
raised an eyebrow	0	Reviewer 2
really cool	12	Reviewer 1
really cool	8	Reviewer 2
red meat	8	Reviewer 1
red meat	12	Reviewer 2
report card	**6**	Reviewer 1
report card	12	Reviewer 2
result in death	**6**	Reviewer 1
result in death	8	Reviewer 2
secret service agents	**6**	Reviewer 1
secret service agents	8	Reviewer 2
send troops	12	Reviewer 1
send troops	8	Reviewer 2
senior citizens	**6**	Reviewer 1
senior citizens	8	Reviewer 2
set up shop	**4**	Reviewer 1
set up shop	**6**	Reviewer 2
sexual activity	**6**	Reviewer 1
sexual activity	8	Reviewer 2
sexually active	**6**	Reviewer 1
sexually active	8	Reviewer 2
short attention span	**6**	Reviewer 1
short attention span	8	Reviewer 2
signed a bill	**6**	Reviewer 1
signed a bill	**4**	Reviewer 2
somehow get	12	Reviewer 1
somehow get	8	Reviewer 2
special education programs	**6**	Reviewer 1
special education programs	12	Reviewer 2
stand ready	8	Reviewer 1
stand ready	**6**	Reviewer 2
still ahead	12	Reviewer 1
still ahead	8	Reviewer 2
stuffed animals	**6**	Reviewer 1
stuffed animals	**4**	Reviewer 2
suddenly found himself	8	Reviewer 1
suddenly found himself	**6**	Reviewer 2
suffered a heart attack	**6**	Reviewer 1
suffered a heart attack	12	Reviewer 2

suggested retail price	**6**	Reviewer 1
suggested retail price	8	Reviewer 2
take a hit	12	Reviewer 1
take a hit	**6**	Reviewer 2
take forever	12	Reviewer 1
take forever	8	Reviewer 2
the ballot box	8	Reviewer 1
the ballot box	12	Reviewer 2
the big bang	**6**	Reviewer 1
the big bang	**4**	Reviewer 2
the black box	**6**	Reviewer 1
the black box	**4**	Reviewer 2
the corner of my eye	12	Reviewer 1
the corner of my eye	**6**	Reviewer 2
the course of history	**6**	Reviewer 1
the course of history	8	Reviewer 2
the dance floor	**6**	Reviewer 1
the dance floor	12	Reviewer 2
the death penalty	**6**	Reviewer 1
the death penalty	8	Reviewer 2
the dress code	**6**	Reviewer 1
the dress code	8	Reviewer 2
the floor plan	**6**	Reviewer 1
the floor plan	8	Reviewer 2
the hardest part of	12	Reviewer 1
the hardest part of	**6**	Reviewer 2
the kind of guy	12	Reviewer 1
the kind of guy	8	Reviewer 2
the medal of honor	**6**	Reviewer 1
the medal of honor	12	Reviewer 2
the middle class	12	Reviewer 1
the middle class	8	Reviewer 2
the old guard	8	Reviewer 1
the old guard	**6**	Reviewer 2
the opposition party	**6**	Reviewer 1
the opposition party	8	Reviewer 2
the other half	12	Reviewer 1
the other half	**6**	Reviewer 2
the party leadership	**6**	Reviewer 1
the party leadership	8	Reviewer 2
the performing arts	**6**	Reviewer 1
the performing arts	8	Reviewer 2
the political spectrum	**6**	Reviewer 1
the political spectrum	8	Reviewer 2
the political will to	**6**	Reviewer 1
the political will to	8	Reviewer 2
the present day	**6**	Reviewer 1
the present day	8	Reviewer 2
the question how	12	Reviewer 1
the question how	8	Reviewer 2

the real deal	8	Reviewer 1
the real deal	0	Reviewer 2
the ruling party	**6**	Reviewer 1
the ruling party	8	Reviewer 2
the scientific community	8	Reviewer 1
the scientific community	12	Reviewer 2
the setting sun	**6**	Reviewer 1
the setting sun	12	Reviewer 2
the space program	**6**	Reviewer 1
the space program	12	Reviewer 2
the visual arts	8	Reviewer 1
the visual arts	12	Reviewer 2
the working class	12	Reviewer 1
the working class	8	Reviewer 2
this year alone	12	Reviewer 1
this year alone	8	Reviewer 2
to address the problem	12	Reviewer 1
to address the problem	**6**	Reviewer 2
to change course	8	Reviewer 1
to change course	**4**	Reviewer 2
to cut down on	8	Reviewer 1
to cut down on	**4**	Reviewer 2
to face reality	12	Reviewer 1
to face reality	8	Reviewer 2
to fight fire with fire	**4**	Reviewer 1
to fight fire with fire	0	Reviewer 2
to find out	8	Reviewer 1
to find out	**6**	Reviewer 2
to fly off	8	Reviewer 1
to fly off	**6**	Reviewer 2
to gain power	12	Reviewer 1
to gain power	8	Reviewer 2
to get a taste of	8	Reviewer 1
to get a taste of	**6**	Reviewer 2
to get comfortable with	12	Reviewer 1
to get comfortable with	8	Reviewer 2
to get hold of	8	Reviewer 1
to get hold of	**6**	Reviewer 2
to get the message	12	Reviewer 1
to get the message	**6**	Reviewer 2
to go along with	12	Reviewer 1
to go along with	**6**	Reviewer 2
to go anyway	12	Reviewer 1
to go anyway	8	Reviewer 2
to go crazy	8	Reviewer 1
to go crazy	**6**	Reviewer 2
to go through	12	Reviewer 1
to go through	8	Reviewer 2
to go to the bathroom	12	Reviewer 1
to go to the bathroom	**6**	Reviewer 2

to grow old	**6**	Reviewer 1
to grow old	12	Reviewer 2
to make a buck	8	Reviewer 1
to make a buck	**4**	Reviewer 2
to pass on the	8	Reviewer 1
to pass on the	0	Reviewer 2
to play ball	**6**	Reviewer 1
to play ball	12	Reviewer 2
to speak out	**6**	Reviewer 1
to speak out	8	Reviewer 2
to turn a profit	8	Reviewer 1
to turn a profit	**6**	Reviewer 2
took a deep breath	8	Reviewer 1
took a deep breath	12	Reviewer 2
trailed off	8	Reviewer 1
trailed off	**4**	Reviewer 2
turned over to	12	Reviewer 1
turned over to	8	Reviewer 2
universal health care	12	Reviewer 1
universal health care	8	Reviewer 2
utility companies	**6**	Reviewer 1
utility companies	8	Reviewer 2
venture capital	**6**	Reviewer 1
venture capital	8	Reviewer 2
wage workers	12	Reviewer 1
wage workers	**4**	Reviewer 2
wake up in the morning	8	Reviewer 1
wake up in the morning	12	Reviewer 2
we simply cannot	8	Reviewer 1
we simply cannot	12	Reviewer 2
well aware of the	8	Reviewer 1
well aware of the	12	Reviewer 2
where the hell is	8	Reviewer 1
where the hell is	**6**	Reviewer 2
why the hell	8	Reviewer 1
why the hell	**6**	Reviewer 2
widely held belief that	8	Reviewer 1
widely held belief that	**6**	Reviewer 2
write a check	**6**	Reviewer 1
write a check	8	Reviewer 2
writer based in	**6**	Reviewer 1
writer based in	8	Reviewer 2
you can possibly	12	Reviewer 1
you can possibly	8	Reviewer 2

Appendix 22. L1-L2 congruency ratings for 11,208 MWUs
(This is just a sample. Full and more detailed data available upon request)

RATING	MWU	MWU JAPANESE TRANSLATION
0	a balancing act	両立が難しい物事
0	a business card	名刺
0	a couple of things	二、三点
0	a critical point	（病気などの）峠
0	a dead end	行き止まり
0	a down payment on	～の頭金
0	a fine line between	～と～は紙一重
0	a fish out of water	場違いの人
0	a green card	永住ビザ
0	a head start on	～の先取り
3	a bad one	良くないもの
3	a bear market	下げ相場
3	a card game	トランプ
3	a convenience store	コンビニ
3	a criminal case	刑事事件
3	a department store	デパート
3	a foster home	児童養護施設
3	a great distance	遠路
3	a hard time	つらい状況
3	a heat wave	猛暑
6	10 minutes or so	10分ほど
6	15 months in prison	懲役15ヶ月
6	2 hours to get	２時間かかる
6	25 years in prison	懲役25年
6	30 minutes set aside	30分を確保する
6	6 months or so	6ヶ月ほど
6	a bad idea	まずい考え
6	a better way to	より良い方法
6	a big deal	大したこと
6	a big smile on his face	満面の笑み
9	12 percent growth in	～の12パーセントの成長
9	3 years in a row	3年連続
9	5 feet of water	水深5フィート
9	a bad feeling	嫌な予感
9	a better job	より良い職業
9	a big surprise	大きな驚き
9	a brick wall	レンガの壁
9	a bright day	良く晴れた日
9	a career choice	職業選択
9	a century ago	1世紀前
12	1 inch thick	1インチの太さ
12	10 percent annual	10パーセントの年間～

12	10 percent annually	年間10パーセント
12	10 percent of adults	10パーセントの大人
12	10 percent of the total	全体の10パーセント
12	10 percent reduction	10パーセントの削減
12	10 percent unemployment	失業率10%
12	10 times a month	月に10回
12	15 hours a day	一日15時間
12	1st grade teacher	一年生の先生

Appendix 23. 3,414 MWUs with an L1-L2 congruency rating of 6 or less
(This is just a sample. Full and more detailed data available upon request)

RATING	MWU	MWU JAPANESE TRANSLATION
6	10 minutes or so	10分ほど
6	15 months in prison	懲役15ヶ月
6	2 hours to get	２時間かかる
6	25 years in prison	懲役25年
6	30 minutes set aside	30分を確保する
6	6 months or so	6ヶ月ほど
6	a bad idea	まずい考え
6	a better way to	より良い方法
6	a big deal	大したこと
6	a big smile on his face	満面の笑み
6	a bit more	もう少しの～
3	a bad one	良くないもの
3	a bear market	下げ相場
3	a card game	トランプ
3	a convenience store	コンビニ
3	a criminal case	刑事事件
3	a department store	デパート
3	a foster home	児童養護施設
3	a great distance	遠路
3	a hard time	つらい状況
3	a heat wave	猛暑
0	a business card	名刺
0	a couple of things	二、三点
0	a critical point	（病気などの）峠
0	a dead end	行き止まり
0	a down payment on	～の頭金
0	a fine line between	～と～は紙一重
0	a fish out of water	場違いの人
0	a green card	永住ビザ
0	a head start on	～の先取り
0	a health club	スポーツジム

Appendix 24. Example sentences created by native speakers for all 11,208 MWUs
(This is just a sample. Full and more detailed data available upon request)

MWU	EXAMPLE SENTENCE
don't know	I don't know what to do, so I will ask my boss.
I don't think	I don't think I will go abroad because I don't have any money.
I don't want to	I don't want to do anything wrong.
had never	Before going to the all-you-can-eat restaurant last night, I had never eaten so much.
how do you	How do you build such incredible ice statues?
years has	The criminal's technique in recent years has been to target retired people.
2 years ago	I graduated from college over 2 years ago, but I still can't find a job.
I will have	I will have a few days off from work next week. Wanna go to the beach?
why do	Why do you always say the same odd things?
all right	Everything is all right. Let's go home.
as well	My sister is going. I'd like to go as well, but I have to work.
know how to	It takes a lot of know how to become a good doctor. There are many things to learn.
high school	I played a lot of sports when I was in high school.
to pick up	Do you think you can drive to the airport to pick up your cousin? I have work so I can't go.
come up with	She is so creative. She can always come up with very clever solutions to tough problems.
had already	I had already decided I wanted to become a doctor before I graduated from high school.

to go back to	After I get out of the Army, I want to go back to school and study law.
come back	I'll probably come back home first, change my clothes, and then go to the party.
has ever	No one has ever passed her test.
can see	I can see a wild monkey over there in the woods.
doesn't mean	Tonight's desert doesn't mean that you are forgiven.
said it will	It said it will rain today, but I don't know...it doesn't look like it will.
to go out	You ought to go out and find someone to marry. Try attending church!
don't really	You don't really think I'm going to give you my salary, do you?
you can get	You can get really great bread from that bakery.
I have heard	I have heard that one out of two marriages fail.
I don't like	I don't like to do such hard work.
don't need	You're fired, so naturally you don't need to come to work tomorrow.
I thought I would	I thought I would be retired at this point in my life, but my savings just weren't enough.
out there	There are many opportunities out there for ambitious people.
how can I	How can I get some service around here? I've been waiting forever to order.
come out	As soon as you come out of the show, call me and I'll come pick you up.
no longer	I am no longer a member of the university soccer team. I quit last week.
get up	I totally couldn't get up this morning, and ended up being 45 minutes late for work.

to find out	My father is dying to find out what I want to do in the future.
they have left	They have left the matter up to me, and I have to decide soon.
I could see	I could see the shore in the distance and realized the cruise would be over soon.
make sure that	Make sure that your seatbelt is on before I start driving the car.
get out of the	I saw them get out of the taxi together.
of health care	The lack of health care for poor people is horrible.
it will take	By train it will take at least an hour, but if you drive it'll only take 30 minutes.
so far he	So far he has only read one book but he will finish three more this year.
grew up	I've been living on the East coast for nearly 20 years, but I actually grew up on the West coast.
men and women	According to recent studies, the brains of men and women are different. Maybe that's why they are always arguing with each other.
to let go	My teacher told me to let go of my fears.
don't feel	I'm tired so I don't feel like going to class.
come in	When you come in to the house, please take off your shoes.
I would like to	I'm writing to you today because I would like to apologize for the way I acted the other day.
turned out to	The cancer in John's body turned out to be much greater than the doctor originally thought.
think so	I don't think it will rain tomorrow, but if you think so you should bring an umbrella.
16 years old	When I was 16 years old, I was able to get my driver's license.
make up	Don't tell your boss the truth about why you're late. Just make up something.

decision making	The decision making at my company is made by both managers and the factory workers.
I want to know	I want to know more about this artist. Can you recommend any books about her?
have lost their	The team may have lost their hunger for victory.
points out that	In his book, the author points out that pollution in this area is getting worse.
what happened to	I'm reading a book about what happened to various child stars.
took place	Yesterday, a meeting took place in the main office between myself and the other staff members.
set up	The company plans to set up a way for customers to check on their order status on its website.
I don't believe	I don't believe in Santa Claus anymore.
I had no idea	I had no idea that he was such a famous star until I looked him up on the Internet.
I think it will	I think it will rain pretty soon. Look at those clouds.
for a long time	I've been working here for a long time. Let's see, it's probably nearly 20 years now.
ask questions	At the end of the speech, we'll have time if anyone wants to ask questions.
people think	At work, people think I'm conservative. They have no idea about the wild things I do on the weekends.
to get back to	I'm eager to get back to my regular life after travelling around Asia.
give up	When I was little, my dad would never let me give up on anything and that had a major impact on who I am today.
I have learned	I have learned to never ask him about politics or religion.
looked up	When I looked up at the sky, I was amazed. There wasn't a cloud in it.
go there	I'd had no desire to go there until he suggested it.

time spent on	My coach said I need to increase my time spent on stretching to avoid getting injured again.
over the past 2 years	We've been trying to renovate our home over the past 2 years, but there's still a lot of work to do.
to go down	I'm expecting the price to go down soon, so I'll buy it later.
end up	My coach said that if I keep exercising like that, I could end up injuring myself, so I changed my workout.
well-known	This restaurant is very well-known for its use of fresh vegetables.
will know	Do you have any idea when you will know if my application has been approved or not?
you can tell	You can tell that he really likes her. Just look at the way he talks to her.
come here	I told the electrician if he could come here in the afternoon it would be best.
people can	People can communicate with each other much easier nowadays because of the internet.
2 years later	We met online, and 2 years later we were married.
even more	I like freshly baked pizza, but I like cold day-old pizza even more.
a lot of people who	A lot of people who live in the city would like to see more bicycles and less cars in the downtown core.
other people	I like Japanese food, and I notice that many other people like it too.
the young man	The young man was driving much too fast. Perhaps he will learn to slow down as he gets older.
take care of	Can you take care of my pet dog while I am away on vacation?
to get there	There are only two ways to get there: car and bus.
sat down	My grandpa sat down on his favorite chair.
take time	To develop a skill like that, it is going to take time so you have to be dedicated if you want to be successful.

how could you	I can't believe you did that! That is so mean! How could you???
opened the door	After I knocked twice, she opened the door.
where do	Where do you think Bob went last night?
figure out	I can't figure out this mathematics problem. It's too difficult.
make it more	The teacher's lesson was very boring, so she tried to make it more interesting.
come on	Come on, George! We're going to be late. Walk faster.
at a time when	I met her at a time when everything was crazy in my life, but after we started dating everything started to get better.
on the other hand	I would love to get married, but on the other hand I enjoy the freedom of being single.
well, I think	Well, I think the economy will eventually recover so now is a great time to invest.
become more	I would like to become more popular because I don't have many friends.
play a role	Scientists say that wind power can play a role in cutting down on pollution.
don't understand	I don't understand why she walks in the rain without an umbrella.

Appendix 25. 50 question collocational fluency test and relevant data
(More detailed data available upon request)

TEST QUESTION

Between work and my kids, finding free time has really become a b _ _ _ _ _ _ _ act.
I was so nervous before my game because if I m _ _ _ _ _ up my coach would be very disappointed.
He is from France. To be more p _ _ _ _ _ _, he is from Paris.
The teacher took r _ _ _ call, and was surprised at how many students didn't come to class that day.
I s _ _ _ _ _ _ _ agree with his views on spending, so I'll probably vote for him.
I will try to get c _ _ _ _ _ _ _ _ _ _ with my computer's new operating system, which is difficult.
The team's victory was d _ _ in part to good coaching.
H _ _ _ is a guy who can give you good advice about cameras.
Now that the war is over, the president has announced his plan to bring the t _ _ _ _ _ home.
I doubt my son is going to follow t _ _ _ _ _ _ on his promise to cut the grass.
I would very much a _ _ _ _ _ _ _ _ _ your help if you have the time.
I was studying for my test all night and it was difficult to stay a _ _ _ _ the next day.
The latest movie was a real winner at the b _ _ office. The company made a lot of money.
The bank offered a line of c _ _ _ _ _ to the company to buy some new equipment.
The boss set the t _ _ _ for the other workers. He is a very hard worker.
I served on the b _ _ _ _ of directors for twenty years.
The dog k _ _ _ of looked sad as it was slowly walking along the side of the road.
You'd think that this evidence would presu _ _ _ _ _ be the one that would send him to jail, but it wasn't.
That s _ _ _ _ _ good. Let's do it!
I don't need my car keys because I'm going to w _ _ _ to the library.
Try to finish as f _ _ _ as you can. We only have a few hours left to finish.
It is said that he was the g _ _ _ _ _ _ _ man that ever lived.
It's so _ _ _ _ _ _ _ hard to find a taxi at rush hour.
When I was a little boy, I had no d _ _ _ _ in my mind that Santa Claus was real.
I was p _ _ _ _ _ _ to see that you did well on your final exam.
You won't m _ _ _ if I borrow this shirt, right?
For testing p _ _ _ _ _ _ _ only, we used the new medicine on volunteers.
It won't be e _ _ _, but in the end you'll agree that the hard work was worth the effort.
If the teacher says to study day and night that doesn't n _ _ _ _ _ _ _ _ _ _ mean go without sleep.
The chances are slim, but it cannot be r _ _ _ _ out.
Nearly one million children died of d _ _ _ _ _ _ in that country last year.
A major f _ _ _ _ _ in the winning of the election was the politician's willingness to be open and honest.
After looking at many health i _ _ _ _ _ _ _ _ plans, we decided to choose this one.
The shopping center has a sp _ _ _ _ _ _ goods store and a shoe store.
The country will v _ _ _ for their new prime minister this month.
Oh my God! Look at that h _ _ _ guy. He looks like he could lift 1,000 pounds.
Refugee camps in n _ _ _ _ _ _ _ _ _ _ countries became overwhelmed as the civil war became worse.
After the robbery, the police officer took a suspect to the police s _ _ _ _ _ _.
In a S _ _ _ _ _ _ Court case the judge may often makes judgments that are referenced later.
It would s _ _ _ _ _ _ _ me if the house didn't sell within a few weeks. It's priced pretty low.
Are you sure that we are going in the right d _ _ _ _ _ _ _ _? I think we're lost.
They stood at the e _ _ _ of the cliff and enjoyed the view.
At the b _ _ _ _ _ _ _ _ of the 19th century, the city was just starting to grow.
Su _ _ _ _ research often uses data collected from a large number of people.
I called his name and he s _ _ _ _ his head out of the passenger window.

The court will hear a _ _ _ _ _ _ _ _ against the new regulations.
I would like to visit the nation's c _ _ _ _ _ _ because there are many famous buildings there.
This is a good e _ _ _ _ _ _ of African art.
Students seem to prefer le _ _ _ _ _ _ activities that are involved with computers.
The movie is about time t _ _ _ _ _. It showed a future where computers control nearly everything.

Appendix 26. Collocational fluency test results
(More detailed data available upon request)

TEACHER	CLASS	TOEFL SCORE	STUDENT ID #	PERCENTAGE CORRECT
Teacher A	Class 1	450	136750	0
Teacher B	Class 1	370	156412	0
Teacher B	Class 2	360	156669	0
Teacher D	Class 1	310	156025	0
Teacher D	Class 2	346	156049	0
Teacher E	Class 3	420	156347	0
Teacher G	Class 2	450	136640	0
Teacher H	Class 3	340	156764	0
Teacher H	Class 3	380	156739	0
Teacher H	Class 3	420	156177	0
Teacher J	Class 4	425	137069	0
Teacher L	Class 2	400	136304	0
Teacher M	Class 1	370	156788	0
Teacher M	Class 1	400	156620	0
Teacher B	Class 1	380	156670	2
Teacher B	Class 1	400	156551	2
Teacher B	Class 1	400	156440	2
Teacher B	Class 1	413	156125	2
Teacher B	Class 2	340	156258	2
Teacher B	Class 2	380	156888	2
Teacher B	Class 2	410	156828	2
Teacher B	Class 2	420	156337	2
Teacher B	Class 3	370	156625	2
Teacher B	Class 3	400	156756	2
Teacher B	Class 3	410	156060	2
Teacher B	Class 3	410	156870	2
Teacher B	Class 3	413	156448	2
Teacher B	Class 3	413	156803	2
Teacher D	Class 1	400	156170	2
Teacher D	Class 2	350	156558	2
Teacher D	Class 2	378	156395	2
Teacher E	Class 1	407	156153	2
Teacher E	Class 2	400	156029	2
Teacher E	Class 3	430	156241	2
Teacher G	Class 1	400	146036	2

Teacher G	Class 2	450	146203	2
Teacher G	Class 3	400	146036	2
Teacher H	Class 1	380	156335	2
Teacher H	Class 1	400	156046	2
Teacher H	Class 2	310	156447	2
Teacher H	Class 2	380	156468	2
Teacher H	Class 3	370	146489	2
Teacher H	Class 3	377	156250	2
Teacher H	Class 3	386	156673	2
Teacher H	Class 3	390	156725	2
Teacher H	Class 3	417	156858	2
Teacher I	Class 1	430	126530	2
Teacher J	Class 1	440	137051	2
Teacher J	Class 4	434	137057	2
Teacher J	Class 4	460	127007	2
Teacher J	Class 5	440	137090	2
Teacher K	Class 1	400	156129	2
Teacher L	Class 1	390	137035	2
Teacher L	Class 1	400	137073	2
Teacher L	Class 3	405	136865	2
Teacher M	Class 1	360	156139	2
Teacher M	Class 1	360	156456	2
Teacher M	Class 1	400	156600	2
Teacher M	Class 1	413	156101	2
Teacher A	Class 1	380	136541	4
Teacher B	Class 1	310	156182	4
Teacher B	Class 1	390	156812	4
Teacher B	Class 2	370	156697	4
Teacher B	Class 2	390	156709	4
Teacher B	Class 2	400	156022	4
Teacher B	Class 3	390	156563	4
Teacher B	Class 3	400	156854	4
Teacher B	Class 3	400	156771	4
Teacher B	Class 3	400	156294	4
Teacher B	Class 3	410	156664	4
Teacher B	Class 3	410	156123	4
Teacher B	Class 3	420	156919	4
Teacher B	Class 3	420	156720	4
Teacher B	Class 3	440	156861	4
Teacher C	Class 1	410	137068	4
Teacher D	Class 1	340	156086	4

Teacher D	Class 1	360	156099	4
Teacher D	Class 1	380	156159	4
Teacher D	Class 1	410	156721	4
Teacher D	Class 2	378	156453	4
Teacher D	Class 2	400	156031	4
Teacher D	Class 2	410	156628	4
Teacher E	Class 1	390	156029	4
Teacher E	Class 1	410	156189	4
Teacher E	Class 1	447	156795	4
Teacher E	Class 2	350	156543	4
Teacher E	Class 2	400	156296	4
Teacher E	Class 3	370	156608	4
Teacher G	Class 1	440	146752	4
Teacher G	Class 1	450	146002	4
Teacher G	Class 2	470	136049	4
Teacher G	Class 3	440	146752	4
Teacher G	Class 3	450	146002	4
Teacher H	Class 1	427	156352	4
Teacher H	Class 1	440	156662	4
Teacher H	Class 2	310	156256	4
Teacher H	Class 3	327	156780	4
Teacher H	Class 3	370	156244	4
Teacher H	Class 3	380	156446	4
Teacher H	Class 3	400	156229	4
Teacher H	Class 3	403	156012	4
Teacher J	Class 1	380	136808	4
Teacher J	Class 1	410	137068	4
Teacher J	Class 2	457	137063	4
Teacher K	Class 1	400	156870	4
Teacher K	Class 1	400	156699	4
Teacher K	Class 1	400	156123	4
Teacher K	Class 1	400	156756	4
Teacher K	Class 1	410	156664	4
Teacher K	Class 1	413	156448	4
Teacher K	Class 1	420	156642	4
Teacher L	Class 1	420	137001	4
Teacher L	Class 3	400	136174	4
Teacher L	Class 3	410	136047	4
Teacher M	Class 1	310	156393	4
Teacher M	Class 1	390	156227	4
Teacher M	Class 1	390	156658	4

Teacher M	Class 1	393	156651	4
Teacher M	Class 1	397	156495	4
Teacher M	Class 1	400	156710	4
Teacher M	Class 1	403	156118	4
Teacher A	Class 1	380	136675	6
Teacher A	Class 1	400	136528	6
Teacher A	Class 1	400	136648	6
Teacher A	Class 2	365	156589	6
Teacher B	Class 1	310	156535	6
Teacher B	Class 1	310	156633	6
Teacher B	Class 1	370	156716	6
Teacher B	Class 1	380	156508	6
Teacher B	Class 1	380	156082	6
Teacher B	Class 1	400	156329	6
Teacher B	Class 1	432	156493	6
Teacher B	Class 2	380	156055	6
Teacher B	Class 2	392	156544	6
Teacher B	Class 2	411	156473	6
Teacher B	Class 3	350	156128	6
Teacher B	Class 3	430	156538	6
Teacher B	Class 3	440	156333	6
Teacher C	Class 1	400	137034	6
Teacher C	Class 1	450	136271	6
Teacher C	Class 1	480	136592	6
Teacher D	Class 1	350	156905	6
Teacher D	Class 1	380	156053	6
Teacher D	Class 1	400	156779	6
Teacher D	Class 1	420	156288	6
Teacher D	Class 2	440	156519	6
Teacher E	Class 1	400	156718	6
Teacher E	Class 1	407	156909	6
Teacher E	Class 1	420	156261	6
Teacher E	Class 1	430	156578	6
Teacher E	Class 2	400	156703	6
Teacher E	Class 2	400	156178	6
Teacher E	Class 3	473	156103	6
Teacher E	Class 4	320	156002	6
Teacher E	Class 4	370	156598	6
Teacher E	Class 4	380	156712	6
Teacher E	Class 4	400	156236	6
Teacher E	Class 4	430	156656	6

Teacher F	Class 2	350	146386	6
Teacher G	Class 2	447	146197	6
Teacher H	Class 1	350	156346	6
Teacher H	Class 1	350	156376	6
Teacher H	Class 1	375	156553	6
Teacher H	Class 1	390	156728	6
Teacher H	Class 1	400	156637	6
Teacher H	Class 2	350	165474	6
Teacher H	Class 2	380	156096	6
Teacher H	Class 2	400	156845	6
Teacher H	Class 2	400	156038	6
Teacher H	Class 3	360	156006	6
Teacher H	Class 3	400	146482	6
Teacher H	Class 3	420	156877	6
Teacher H	Class 3	433	156104	6
Teacher I	Class 1	362	136501	6
Teacher I	Class 1	500	136052	6
Teacher J	Class 2	420	137089	6
Teacher J	Class 3	450	136665	6
Teacher J	Class 4	420	127002	6
Teacher J	Class 4	440	127020	6
Teacher J	Class 4	450	126130	6
Teacher J	Class 5	430	136789	6
Teacher K	Class 1	403	156320	6
Teacher K	Class 1	420	156720	6
Teacher L	Class 1	380	137102	6
Teacher L	Class 2	417	116523	6
Teacher L	Class 3	423	136065	6
Teacher L	Class 3	437	136172	6
Teacher L	Class 3	440	126204	6
Teacher M	Class 1	373	156809	6
Teacher M	Class 1	390	156871	6
Teacher A	Class 1	390	136531	8
Teacher A	Class 1	440	136081	8
Teacher A	Class 2	460	156235	8
Teacher B	Class 1	350	156824	8
Teacher B	Class 1	380	156667	8
Teacher B	Class 1	382	156819	8
Teacher B	Class 1	400	156899	8
Teacher B	Class 2	385	156452	8
Teacher B	Class 3	420	156556	8

Teacher B	Class 3	443	156475	8
Teacher C	Class 1	455	136873	8
Teacher C	Class 1	470	136804	8
Teacher D	Class 1	310	156187	8
Teacher D	Class 1	350	156092	8
Teacher D	Class 1	400	156218	8
Teacher D	Class 1	418	156063	8
Teacher D	Class 1	420	156107	8
Teacher D	Class 2	400	156434	8
Teacher D	Class 2	420	156313	8
Teacher D	Class 2	420	156822	8
Teacher D	Class 2	467	156303	8
Teacher E	Class 1	386	156772	8
Teacher E	Class 1	390	156361	8
Teacher E	Class 1	400	156530	8
Teacher E	Class 1	410	156778	8
Teacher E	Class 1	440	156719	8
Teacher E	Class 2	370	156059	8
Teacher E	Class 2	390	156541	8
Teacher E	Class 3	390	156765	8
Teacher E	Class 3	400	156694	8
Teacher E	Class 4	380	156027	8
Teacher E	Class 4	383	156209	8
Teacher E	Class 4	410	156813	8
Teacher E	Class 4	420	156094	8
Teacher F	Class 2	435	146447	8
Teacher G	Class 1	453	146620	8
Teacher G	Class 2	420	146310	8
Teacher G	Class 2	430	136787	8
Teacher G	Class 2	435	137057	8
Teacher G	Class 3	453	146620	8
Teacher H	Class 1	420	156173	8
Teacher H	Class 2	390	156748	8
Teacher H	Class 2	400	156071	8
Teacher H	Class 2	403	156093	8
Teacher H	Class 2	412	156245	8
Teacher H	Class 3	390	156485	8
Teacher H	Class 3	420	156713	8
Teacher H	Class 3	480	156889	8
Teacher I	Class 1	350	136061	8
Teacher I	Class 1	400	126451	8

Teacher I	Class 1	400	137034	8
Teacher I	Class 1	400	136298	8
Teacher I	Class 1	410	137086	8
Teacher I	Class 1	450	137087	8
Teacher J	Class 1	400	137034	8
Teacher J	Class 1	470	137027	8
Teacher J	Class 2	350	137010	8
Teacher J	Class 2	400	136394	8
Teacher J	Class 2	440	136190	8
Teacher J	Class 4	450	127039	8
Teacher J	Class 5	350	126792	8
Teacher K	Class 1	390	156861	8
Teacher K	Class 1	400	156771	8
Teacher K	Class 1	410	156060	8
Teacher K	Class 1	410	156531	8
Teacher K	Class 1	420	156919	8
Teacher K	Class 1	430	156538	8
Teacher K	Class 2	400	156587	8
Teacher L	Class 1	390	137097	8
Teacher L	Class 1	437	137059	8
Teacher L	Class 2	360	126470	8
Teacher L	Class 2	430	136770	8
Teacher L	Class 3	440	136053	8
Teacher M	Class 1	340	156815	8
Teacher M	Class 1	360	156268	8
Teacher M	Class 1	400	136764	8
Teacher A	Class 1	400	127019	10
Teacher A	Class 1	430	127015	10
Teacher A	Class 2	430	156254	10
Teacher B	Class 1	310	156200	10
Teacher B	Class 1	380	156700	10
Teacher B	Class 2	370	156746	10
Teacher B	Class 2	394	156344	10
Teacher C	Class 1	473	136847	10
Teacher C	Class 1	477	136054	10
Teacher D	Class 1	370	156597	10
Teacher D	Class 2	370	156330	10
Teacher D	Class 2	405	156701	10
Teacher D	Class 2	410	156270	10
Teacher D	Class 2	430	156277	10
Teacher E	Class 1	377	156761	10

Teacher E	Class 2	310	156180	10
Teacher E	Class 3	417	156391	10
Teacher E	Class 3	432	156692	10
Teacher E	Class 3	450	156685	10
Teacher E	Class 3	480	156411	10
Teacher E	Class 4	380	156198	10
Teacher E	Class 4	380	156459	10
Teacher E	Class 4	420	156020	10
Teacher E	Class 4	423	156801	10
Teacher E	Class 4	425	156298	10
Teacher G	Class 1	460	146758	10
Teacher G	Class 3	460	146758	10
Teacher H	Class 1	403	156488	10
Teacher H	Class 2	400	156612	10
Teacher H	Class 2	400	156833	10
Teacher H	Class 2	410	156653	10
Teacher H	Class 3	400	156205	10
Teacher H	Class 3	402	156041	10
Teacher H	Class 3	427	156078	10
Teacher H	Class 3	430	156043	10
Teacher H	Class 3	434	156343	10
Teacher I	Class 1	450	120208	10
Teacher J	Class 1	390	137091	10
Teacher J	Class 1	390	137074	10
Teacher J	Class 4	430	127032	10
Teacher K	Class 1	350	156128	10
Teacher K	Class 1	370	156625	10
Teacher K	Class 1	410	156563	10
Teacher K	Class 1	410	156705	10
Teacher K	Class 1	420	156556	10
Teacher K	Class 1	440	156333	10
Teacher L	Class 1	400	136197	10
Teacher L	Class 2	400	136833	10
Teacher L	Class 3	310	136051	10
Teacher L	Class 3	430	136581	10
Teacher A	Class 1	420	137028	12
Teacher C	Class 1	400	136366	12
Teacher C	Class 1	430	136079	12
Teacher C	Class 1	560	136023	12
Teacher D	Class 1	400	156233	12
Teacher D	Class 1	400	156359	12

Teacher D	Class 1	403	156804	12
Teacher D	Class 1	417	156048	12
Teacher D	Class 2	401	156479	12
Teacher D	Class 2	403	156397	12
Teacher E	Class 1	412	156489	12
Teacher E	Class 1	423	156372	12
Teacher E	Class 1	425	156735	12
Teacher E	Class 2	400	156109	12
Teacher E	Class 2	400	156672	12
Teacher E	Class 3	400	156293	12
Teacher E	Class 3	450	156708	12
Teacher E	Class 3	470	156621	12
Teacher E	Class 4	380	156912	12
Teacher E	Class 4	423	156915	12
Teacher E	Class 4	437	156458	12
Teacher G	Class 1	470	146546	12
Teacher G	Class 1	473	146436	12
Teacher G	Class 3	470	146546	12
Teacher G	Class 3	473	146436	12
Teacher H	Class 1	400	156557	12
Teacher H	Class 2	375	156074	12
Teacher H	Class 2	400	156805	12
Teacher H	Class 2	493	156115	12
Teacher H	Class 3	392	156429	12
Teacher J	Class 1	350	137047	12
Teacher J	Class 2	400	136040	12
Teacher J	Class 2	450	136818	12
Teacher J	Class 4	400	137037	12
Teacher J	Class 4	400	136434	12
Teacher J	Class 4	450	127085	12
Teacher J	Class 5	450	136204	12
Teacher J	Class 5	480	136592	12
Teacher K	Class 1	400	156854	12
Teacher K	Class 2	460	156130	12
Teacher L	Class 1	440	137077	12
Teacher L	Class 3	426	136068	12
Teacher L	Class 3	430	136215	12
Teacher L	Class 3	430	136535	12
Teacher M	Class 1	380	156106	12
Teacher A	Class 1	350	136166	14
Teacher A	Class 1	459	137015	14

Teacher B	Class 1	380	156568	14
Teacher B	Class 2	400	156062	14
Teacher C	Class 1	420	126308	14
Teacher C	Class 1	430	136458	14
Teacher D	Class 1	397	156437	14
Teacher D	Class 1	400	156185	14
Teacher D	Class 1	440	156443	14
Teacher D	Class 2	380	156144	14
Teacher D	Class 2	450	156862	14
Teacher E	Class 1	438	156596	14
Teacher E	Class 1	450	156464	14
Teacher E	Class 2	440	156362	14
Teacher E	Class 3	470	156114	14
Teacher F	Class 1	400	146519	14
Teacher F	Class 1	440	136509	14
Teacher F	Class 1	520	146356	14
Teacher F	Class 2	470	146450	14
Teacher F	Class 2	500	146495	14
Teacher G	Class 1	400	146093	14
Teacher G	Class 3	400	146093	14
Teacher H	Class 1	397	156774	14
Teacher H	Class 1	403	156886	14
Teacher H	Class 2	310	156133	14
Teacher H	Class 2	310	156058	14
Teacher H	Class 2	357	156119	14
Teacher H	Class 2	387	156536	14
Teacher H	Class 3	420	156319	14
Teacher I	Class 1	430	126158	14
Teacher J	Class 1	400	136201	14
Teacher J	Class 1	420	137023	14
Teacher J	Class 2	390	137005	14
Teacher J	Class 2	400	137058	14
Teacher J	Class 2	420	137039	14
Teacher J	Class 3	500	136165	14
Teacher J	Class 4	450	136511	14
Teacher J	Class 5	430	137095	14
Teacher J	Class 5	440	136879	14
Teacher J	Class 5	440	136571	14
Teacher K	Class 2	400	156235	14
Teacher K	Class 2	427	156289	14
Teacher K	Class 2	430	156254	14

Teacher L	Class 1	420	137071	14
Teacher L	Class 1	450	137060	14
Teacher L	Class 3	450	136716	14
Teacher A	Class 1	463	136593	16
Teacher A	Class 1	480	136703	16
Teacher A	Class 2	427	156289	16
Teacher A	Class 2	470	156130	16
Teacher B	Class 1	400	156369	16
Teacher C	Class 1	470	137015	16
Teacher C	Class 1	480	136057	16
Teacher C	Class 1	490	136614	16
Teacher D	Class 1	395	156232	16
Teacher D	Class 1	460	156377	16
Teacher E	Class 1	390	156506	16
Teacher E	Class 4	310	156916	16
Teacher E	Class 4	407	156626	16
Teacher E	Class 4	450	156752	16
Teacher F	Class 1	400	146403	16
Teacher F	Class 1	550	137020	16
Teacher F	Class 2	360	146262	16
Teacher G	Class 2	440	136699	16
Teacher H	Class 1	420	156641	16
Teacher I	Class 1	447	136712	16
Teacher I	Class 1	470	136273	16
Teacher I	Class 1	550	116054	16
Teacher J	Class 1	463	136316	16
Teacher J	Class 2	483	137079	16
Teacher J	Class 3	450	136658	16
Teacher J	Class 3	470	136018	16
Teacher J	Class 3	470	126564	16
Teacher J	Class 3	470	136461	16
Teacher J	Class 4	440	137094	16
Teacher J	Class 5	430	136073	16
Teacher J	Class 5	440	136534	16
Teacher J	Class 5	450	136313	16
Teacher J	Class 5	480	136228	16
Teacher L	Class 2	430	136443	16
Teacher L	Class 2	450	136795	16
Teacher L	Class 3	440	136743	16
Teacher A	Class 1	410	136481	18
Teacher A	Class 1	470	137066	18

Teacher A	Class 1	483	126183	18
Teacher A	Class 2	380	156895	18
Teacher C	Class 1	470	137066	18
Teacher E	Class 1	410	156594	18
Teacher E	Class 1	420	156287	18
Teacher E	Class 3	437	156282	18
Teacher F	Class 1	450	136749	18
Teacher F	Class 1	453	146275	18
Teacher F	Class 1	463	146097	18
Teacher F	Class 2	430	146401	18
Teacher F	Class 2	490	146673	18
Teacher G	Class 2	453	136138	18
Teacher H	Class 1	493	156660	18
Teacher H	Class 2	380	156875	18
Teacher I	Class 1	417	136637	18
Teacher J	Class 3	470	136579	18
Teacher J	Class 3	507	136479	18
Teacher J	Class 4	450	127026	18
Teacher J	Class 4	480	126334	18
Teacher J	Class 4	540	126708	18
Teacher J	Class 5	480	136045	18
Teacher K	Class 2	432	156112	18
Teacher L	Class 2	450	126639	18
Teacher L	Class 3	447	136744	18
Teacher A	Class 2	400	156255	20
Teacher C	Class 1	460	136760	20
Teacher C	Class 1	477	136759	20
Teacher C	Class 1	494	126598	20
Teacher D	Class 2	375	156555	20
Teacher E	Class 3	443	156496	20
Teacher E	Class 3	450	156321	20
Teacher H	Class 1	380	156631	20
Teacher J	Class 3	490	136412	20
Teacher J	Class 5	450	136344	20
Teacher K	Class 2	406	156895	20
Teacher L	Class 1	390	136796	20
Teacher L	Class 1	460	136866	20
Teacher A	Class 1	460	136035	22
Teacher A	Class 1	470	136351	22
Teacher A	Class 1	480	137052	22
Teacher A	Class 1	487	136246	22

Teacher C	Class 1	400	136565	22
Teacher C	Class 1	450	136690	22
Teacher C	Class 1	480	137072	22
Teacher E	Class 3	457	156509	22
Teacher E	Class 4	380	156902	22
Teacher F	Class 2	423	146239	22
Teacher F	Class 2	540	146488	22
Teacher H	Class 1	440	156623	22
Teacher I	Class 1	440	136200	22
Teacher J	Class 4	630	126567	22
Teacher K	Class 2	430	156065	22
Teacher L	Class 1	595	137013	22
Teacher L	Class 2	310	136143	22
Teacher L	Class 3	365	136609	22
Teacher A	Class 1	480	136425	24
Teacher E	Class 3	490	156627	24
Teacher G	Class 2	540	136221	24
Teacher J	Class 2	310	136143	24
Teacher J	Class 3	480	136502	24
Teacher J	Class 4	430	126089	24
Teacher J	Class 4	490	126590	24
Teacher J	Class 5	480	136364	24
Teacher K	Class 2	400	156255	24
Teacher L	Class 3	483	136639	24
Teacher A	Class 2	413	156065	26
Teacher C	Class 1	477	136397	26
Teacher E	Class 3	460	156492	26
Teacher E	Class 3	470	156141	26
Teacher I	Class 1	400	136211	26
Teacher J	Class 3	440	136451	26
Teacher J	Class 3	480	136257	26
Teacher K	Class 2	430	156773	26
Teacher A	Class 1	490	136385	28
Teacher A	Class 2	450	156856	28
Teacher C	Class 1	447	127040	28
Teacher C	Class 1	490	136191	28
Teacher A	Class 1	470	126282	30
Teacher A	Class 1	517	126452	30
Teacher A	Class 2	430	156773	30
Teacher F	Class 2	433	146681	30
Teacher J	Class 3	550	137020	30

Teacher J	Class 5	490	126391	30
Teacher J	Class 5	520	136038	30
Teacher F	Class 1	470	136413	32
Teacher F	Class 1	480	146196	32
Teacher F	Class 2	500	146277	32
Teacher J	Class 3	510	136376	32
Teacher I	Class 1	450	126055	34
Teacher J	Class 3	510	136312	34
Teacher J	Class 5	540	136221	34
Teacher L	Class 3	470	137048	34
Teacher A	Class 1	520	136095	36
Teacher J	Class 5	440	136430	38
Teacher J	Class 5	490	136191	38
Teacher F	Class 1	540	146421	40
Teacher J	Class 3	490	126671	40
Teacher J	Class 3	500	136100	40
Teacher J	Class 3	560	136551	40
Teacher F	Class 1	527	146188	42
Teacher J	Class 5	520	136692	42
Teacher J	Class 4	480	126541	44
Teacher J	Class 5	677	126246	48
Teacher I	Class 1	550	127107	52