N + P CLUSTERS IN FRESHMAN COMPOSITION:

A LEXICO-GRAMMATICAL APPROACH TO ACADEMIC VOCABULARY

FOR SECOND LANGUAGE WRITERS

by

ELIZABETH C. CRAIG

(Under the Direction of Sarah Blackwell)

ABSTRACT

This study addresses the lexical difficulties that English as a second/foreign language learners demonstrate with regard to preposition usage in their academic writing. After taking a look at the types of errors learners make with regard to prepositions, this study examines native speaker usage of N + P clusters in a 500,000-word corpus of freshman essays at a four-year, tier-one research university in the southeastern U.S. N + P clusters designate those nouns that are commonly post-modified by prepositional phrases functioning adjectivally. An N + P cluster then consists of a preposition plus its most frequent and robust nominal left colligates as in *access to*, *amount(s) of*, *increase(s) in,* and *effect(s) on*. N + P clusters used with high frequencies by native speakers in the academic register of expository writing are found with the aid of a concordancer software program by first targeting the ten most frequent prepositions in the Corpus and then determining their most frequent nominal left colligates. The degree of attraction between particular nouns and prepositions is determined through a proportional analysis, and a semantic taxonomy of the most robust N + P clusters is then applied as an aid to

functional presentations of academic vocabulary. It is suggested that the teaching of such N + P clusters in a lexico-grammatical approach would benefit L2 learners in their efforts to achieve native-like fluency and accuracy with regard to preposition usage and nominal density in second language writing. Included are implications for the further investigation of N + P clusters in academic writing for EAP materials design, especially for content-area vocabulary.

INDEX WORDS: Academic vocabulary; Second language writing; Collocations; Corpus linguistics; L2 Prepositions; Nominal density; Lexico-grammatical approach.

N + P CLUSTERS IN FRESHMAN COMPOSITION:
A LEXICO-GRAMMATICAL APPROACH TO ACADEMIC VOCABULARY
FOR SECOND LANGUAGE WRITERS

by

ELIZABETH C. CRAIG

B.A., The University of Georgia, 1981

M.S., Georgia State University, 1994

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2008

N + P CLUSTERS IN FRESHMAN COMPOSITION:

A LEXICO-GRAMMATICAL APPROACH TO ACADEMIC VOCABULARY

FOR SECOND LANGUAGE WRITERS

by

ELIZABETH CLAIBORNE CRAIG

Major Professor: Sarah Blackwell

Committee: William Kretzschmar
Margaret Quesada

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2008

DEDICATION

This dissertation is dedicated to my mother, Betty Anne Oliver Craig, who would have been proud beyond words, and to my father, Raymond Archer Craig, Jr., who would have been tickled pink.

ACKNOWLEDGEMENTS

I would like to thank my Director and editor extraordinaire, Dr. Sarah Blackwell, for making me readable, and my committee members, Dr. Margaret Quesada and Dr. William A. Kretzschmar, Jr., for helping me focus. I would also like to thank the First-Year Composition team at The University of Georgia, Dr. Christy Desmet, Dr. Deborah Church Miller, and Dr. Ron Balthazor, for allowing me the opportunity to conduct this research in the first place. A special thanks goes to Stael Ruffinelli de Ortiz and Juan Antonio Avalos Pinto, who gave me access to their students in South America and to all the students, both near and far, who participated by providing the fodder for this study. It is their linguistic relics without which this analysis would have been impossible. Also, I would like to express my immense gratitude to two people I have never met, Dr. Laurence Anthony (Waseda University, Japan) and Dr. Paul Rayson (Lancaster University, UK), who engineered the software programs utilized for the essential analysis herein. Without their generous and individual assistance, I would have been lost in a sea of data.

Small words make big meanings

The hundred or so short and frequent words of English have two roles in the making of meaning. They sometimes give grammatical information, and so they are allotted to word classes. This tells us little about them as individuals, but it locks them up in the grammar, and we think of nouns, verbs, adjectives and adverbs as the individual members of the vocabulary.

The study of the way words occur, pattern and combine in a large text corpus presents a different picture. Here, small words make big meanings. We must move on from a view of the vocabulary as consisting mainly of single-word items to one where phrase patterns are prominent and insistent. In the phrase patterns, all the constituent words are of equal status, and often it is the small, hardly-noticed words that provide the crucial identification of a meaningful unit.

For someone seeking mastery of a language there is a lot to be gained from working with the actual meaningful units from an early stage, avoiding needless analysis; corpus research, properly focused, can sharpen perceptions of meaning, offer accurate models of usage and speed up learning by concentrating on those patterns which are the most widespread and pervasive – those which involve the small words.

----------John Sinclair
Plenary Address
2006 AAAL Conference
Montreal, Quebec

(Retrieved from http://www.aaal.org/aaal2006/sinclair.htm on November 17, 2008)

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 Problem Statement

This study addresses the lexical difficulties that English as a second/foreign language learners demonstrate with regard to preposition usage in their academic writing. Such small but common function words like articles and prepositions are notoriously difficult for even advanced, non-native speakers (NNSs) of English. Indeed, the absence of articles in second language (L2) English can be particularly distinctive of speakers of Asian and Eastern European languages, which make little or no use of these small words. L2 learners of English also either omit or misuse prepositions because English contains a relatively rich array of them with very fine distinctions in their distribution of use, which can leave the learner to guess at which one to use in many instances, perhaps often relying on direct translation from the first language (L1).

We typically think of prepositions as functioning as parts of prepositional phrases and in phrasal verbs, but they also occur in patterns with particular preceding nouns more often than one might expect. A good example of a phrasal adjective derived from a verb form occurs in the previous sentence (*functioning as*), and a good example of an N + P cluster follows (*parts of*). Academic writing is full of such N + P clusters "because of the frequent need for definition and specification" (Carter & McCarthy, 2006, p. 269) in such formal, informationally-dense registers.

Though they might be dismissed as insignificant, minor, or 'local' errors by some second language writing (SLW) teachers (Ferris, 2002), errors with regard to function words remain an identifying (Benson, Deming, Denzer, & Valeri-Gold, 1992; Reid, 1988) and therefore stigmatizing characteristic of NNS production. For example, upon hearing or reading the phrase *confidence on myself* produced by an adult, the native speaker (NS) of English immediately recognizes a NNS as such. Furthermore, accuracy ensures that intended messages are conveyed.

> Conventionality of style…aids precision of expression, clearly a quality highly valued in academic argument…While the collocational errors they [L2 students] make do not on the whole seriously destroy intelligibility, they can lead to a lack of precision and obscure the clarity required in academic communication. (Howarth, 1996, p. ix)

Some examples of actual L2 errors with regard to prepositions following nouns in the present study include *example for this quality, city from Argentina,* and *a look on the themes* (see Section 3.2).

Prepositional phrases, especially those used to indicate spatial or temporal relationships such as *in*, *at*, and *on*, have been addressed in ESL/EFL teaching materials (for example, see Azar, 2003) and classrooms for quite some time. In addition, there is a plethora of L2 teaching and reference materials on phrasal verbs such as *come in*, *keep on*, and *look over* (for example, see Azar, 2003; Flower, 2002; McCarthy & O'Dell, 2004, 2007)[1] and even some coverage of adjective phrases with prepositions such as *afraid of, interested in, responsible for, anxious about, content with,* and so forth (for example, see Azar, 2003; Cowan, 2008; Raimes, 2004). Yet N + P clusters have been

[1] Also, see the many reference dictionaries for phrasal verbs from ESL publishers, such as Cambridge, Oxford, Heinle, Longman, and Collins.

overlooked entirely as a viable teaching point for L2 English applications. The present study seeks to highlight the common usage of N + P clusters by native speakers (NSs) in their college-level academic writing for the benefit of L2 teachers and students in a lexico-grammatical approach, which has already been exploited with regard to English verbs and adjectives that co-occur with prepositions.

Because of their multiple and abstract meanings, prepositions remain a problematic area in both general linguistics and foreign language education. Prepositions have received less attention than other more semantically-weighty word classes, yet they play a crucial role in mediating between verbs and any nominal objects they may take and in relating noun phrases to each other within sentences. Due to the substantial influence of context on prepositional meaning (or on any word for that matter), students resorting to conventional dictionaries for clarification may become frustrated, or even worse, confused.

> The entries in most dictionaries are indeed not very helpful about words like, *the, of, and* --- the most common words in the language. Because dictionaries traditionally give priority to semantic meaning, as against the meaning found in grammar, usage, and pragmatics, they try to analyse the words by semantic criteria. This is a difficult task, indeed, these very words are frequently said to lack semantic meaning altogether. (Sinclair, 1991b, p. 81)

And Kennedy (2003) agrees that

> part of the learning difficulty of prepositions arises from the fact that most of them have many meanings or uses. The most frequent, *of* and *in*, each have over 40 senses given in comprehensive dictionaries. It is often hard for learners of English to know which preposition to use with particular nouns or verbs…Although prepositions are hard, most courses do not give them enough attention, and learners are often left to learn how to use them as best they can. Too much attention is usually given to literal, physical uses, whereas most prepositions are used with extended meanings that are abstract and figurative. (pp. 251-252)

What follows is a collocational approach to prepositions. For the reader more interested in an exhaustive, semantic description of English prepositions for L2 teachers, see Lindstromberg (1998) and for ESL students, see Yates (1999). For an explanatory discussion on the second language acquisition of certain English prepositions, see Thomas (2004). For a cognitive/semantic approach to teaching L2 prepositions, see Boers and Demecheleer (1998). For a cognitive/semantic treatment of spatial prepositions, see Tyler and Evans (2003). And for a contemporary look at prepositions in their syntactic, semantic, and pragmatic contexts, see Feigenbaum and Kurzon (2002).

1.2 What are N + P Clusters?

The label 'N + P cluster' is meant to refer to two-word phrases involving a noun plus an ensuing preposition and to distinguish this structure from simple noun phrases, which consist of a head noun plus any preceding modifiers such as determiners, adjectives, and other nouns functioning adjectivally as in *the big, yellow school bus*. For the present study, the focus will be on nouns that cluster with an immediately following preposition functioning adjectivally along with its object to somehow modify or clarify the preceding noun. As analogous in structure to prepositional verbs such as *consist of, look at,* and *hope for*, which have received much warranted attention in English language teaching and reference materials to date, N + P clusters are presented here as two-word sequences abundant in formal written registers (see Section 2.4) and consisting of a noun plus its most frequent prepositional post-modifier, as in *access to, amount(s) of,* and *change(s) in*, for exploitation in L2 academic vocabulary presentations.

Hence, the focus of the present study is on prepositions as one of the most frequent and therefore useful parts of speech in written academic English. Prepositions will be targeted as a direct way to find very common content words associated with them, in this case nouns that most often occur in their immediate vicinity and thus may be regarded as forming a cluster along with the attendant preposition. In ESL/EFL language teaching, we have given much attention to the explicit teaching of multi-word verbs as very useful to L2 learners of English because of their ubiquitous nature, especially in conversation (Biber, 1988). One of the outcomes of the present study may be that some concerted effort and attention will be directed at the utility of N + P clusters in informationally-dense writing such as that which we expect from our university inductees in freshman writing courses. "We need to teach basic writers how to manipulate the structures, the syntactic units, not [just] how to identify words" in isolation (Dykstra, 1997, p. 139). Prepositions, in their capacity to provide links among words in a sentence, should be considered quintessential cohesive devices at the phrase level.

N + P clusters are especially suited to a collocational approach because their prepositional components are relatively fixed, whereas adverbial prepositional phrases are highly mobile. Adverbial prepositional phrases can be placed almost anywhere in a clause while maintaining their direct association with the inflected verb such as in *In a little while, he will announce the results*. The introductory prepositional phrase here clearly answers the adverbial question 'when?' about the main verb *announce*. When functioning as adjectivals, however, prepositional phrases tend to remain close to their head nouns, much like relative clauses, in order to avoid potential confusion such as in *In the election, he will announce the results*, where it is not clear whether the prepositional

phrase is functioning adjectivally or adverbially, i.e. modifying *results* or *announce*. Adjectival modifiers are much more restricted in English with regard to movement, whereas adverbials remain the most mobile structures in the language. This relative fixedness for adjectival modifiers suggests a collocational approach, which essentially looks for words that occur together (though not necessarily adjacent to each other) with some regularity.

> In its adjectival role the prepositional phrase identifies the noun headword in relation to time, place, direction, purpose, origin and the like…An adjectival prepositional phrase helps to identify a noun or pronoun by answering the questions 'Which one?' or 'What kind of?' In the case of the adjectival prepositional phrase, we nearly always have a noun phrase within a noun phrase. (Kolln & Funk, 2006, pp. 144-145)

A method based on frequency will serve to highlight those N + P clusters that are relatively more fixed with regard to preposition selection after particular nouns. In other words, in *the boat at the dock*, the preposition selection is more flexible and dependent on the following object, whereas in *the crux of the matter*, the preposition selection is more fixed (idiomatic) and determined by the preceding noun, *crux*.

As for the structure of such complex noun phrases, we can represent the restrictions on word order by type of post-modifier:

> [Noun Phrase] [Prep Phrase] [Participial Phrase] [Relative Clause]
> (*Our access*) (*to the website*) (*having been granted*), (*which was temporary*…)

In other words, if a noun phrase is post-modified by a prepositional phrase, it generally precedes all other types of post-modification (Kolln & Funk, 2006), hence, adjectival prepositional phrases tend to be located right next to their respective nouns. So by looking at the immediate left collocates of prepositions in a corpus, the investigator will

be able to identify the particular nouns that precede and are commonly modified by particular prepositions.

Possibly due to the popularity of syntactic theory in the U.S., we are accustomed to separating noun phrases from their prepositional phrase complements because of a formal rule: NP + PP. It is suggested here that we consider re-analyzing these structures as the collocational patterns (*the crux of*) (*the matter in*) (*this paper*) in order to establish the close association a preposition can have with its preceding noun. A collocational approach takes into consideration both the syntagmatic and paradigmatic axis.

> Whereas *syntax* deals with general classes of words and their combinations, *collocations* describe specific lexical items and the frequency with which these items occur with other lexical items. Collocations are defined along a syntagmatic, or horizontal, dimension and a paradigmatic, or vertical dimension. That is, a collocational unit consists of a 'node' that co-occurs with a span of words on either side. The span consists of particular word classes filled by specific lexical items. (Nattinger & DeCarrico, 1992, p. 20)

Swick (2005, p. 62) identifies the following preposition plus noun compounds: *bylaw, bypass, downfall, infield, insight, outbreak, outgrowth, outline, outlook, underarm, underclassman, underwear, upheaval,* and *uproar*. We may suppose that these compounds were once written variably as two separate words, as hyphenated compounds, or as one word, the latter form having eventually won out, however tenuously, as these forms became regarded as individual semantic units because of strong collocational tendencies. Indeed, one author treats the following structures as single constituents, which he likes to call 'prearticles:' *a little of, plenty of, a lot of, a good deal of, a small quantity of, an item of, a slice of* (Morenberg, 2002, p. 82). Should we not consider the extension of this tendency toward lexicalization to other very common noun-

preposition combinations? The preposition *of* is the most prolific preposition in English and a very common nominal post-modifier; a corpus frequency and collocational analysis will bear this out and perhaps even reveal more such useful little words.

1.3 Treatment of N + P Clusters as Lexical Units

By analogy to prepositional verbs, N + P clusters can be thought of as the simple structure N + P taking a nominal object. However, in their treatment of 'multi-word lexical verbs,' Biber, Johansson, Leech, Conrad, and Finegan (1999) distinguish four types: phrasal verbs, prepositional verbs, phrasal-prepositional verbs, and other multi-word verbs. All four types are described as "relatively idiomatic units" (p. 403) that function as single lexical items. For examples based on usage, the authors provide:

- verb + adverbial particle: phrasal verbs, e.g. *pick up*
- verb + preposition: prepositional verbs, e.g. *look at*
- verb + particle + preposition: phrasal-prepositional verbs, e.g. *get away with*
- other multi-word verb constructions, notably:
  verb + noun phrase (+ preposition) e.g. *take a look (at)*;
  verb + prepositional phrase, e.g. *take into account*;
  verb + verb, e.g. *make do*. (p. 403)

For Biber et al. (1999), the key to drawing this fine a distinction between phrasal and prepositional verbs lies in the consideration that the second element is an adverbial particle in the former, with closer ties to the verb, and a preposition in the latter, requiring an object. However, they go on to say that "in practice, it is hard to make an absolute distinction between free combinations and fixed multi-word verbs; one should rather think of a cline on which some verbs, or uses of verbs, are relatively free and others relatively fixed" (p. 403). From a historical point of view with regard to multi-word

verbs, Brinton and Traugott (2005) argue that the particles of phrasal verbs represent a grammaticalization process and prepositional verbs have been lexicalized (p. 123). A collocational analysis as is undertaken below with regard to N + P clusters can establish the relative strength of such relationships among the two words and serves as an empirical way to determine the phrasal status of these contiguous elements (N + P) rather than relying on any native speaker intuitions, which can be faulty even for language teachers (McCrostie, 2007).

With regard to N + P clusters, the strength of the attraction between the noun and certain prepositions will be established through an examination of proportional distributions in the NS Corpus, thereby eliminating those contenders for N + P cluster status not having a strong enough attraction to warrant their treatment as single lexical units. Certain nouns take certain prepositions in their wake with some regularity, and hence, it would benefit the student to recognize and learn to use them appropriately in their academic writing. Each concordance of a preposition following a noun will also have to be checked individually for potential association with a preceding, separable phrasal verb as with *put in* in *He put many hours in*, where *hours in* would not be a contender for N + P cluster status here because *in* is essentially part of the preceding phrasal verb, in other words, an adverbial particle.

1.4 Background and General Definitions

Prepositions are relatively small and frequent function words used to indicate spatial, temporal, or more abstract relationships among words in a sentence. They can also be thought of as analogous to inflectional suffixes, which present special challenges

for adult learners as well because of their lesser salience in word-final, unstressed position. During the Middle English period of much syntactic change, prepositions won out in the language over many inflectional endings that had existed in Old English and that were redundant to the prepositional functions already at hand. "Prepositions like *in, with,* and *by* came to be used more frequently than in Old English" (Barber, 1993). Both prepositions and the few remaining inflectional suffixes in English serve to tie words to each other in a meaningful way in sentences. In the British tradition, Firth's (1957) 'contextual theory of meaning,' which considers a word's collocations as an intrinsic part of its meaning, Halliday's (1991) 'probabilistic grammar,' and Sinclair's (1991b) corpus-informed language teaching are guiding principles for this collocational analysis of prepositions as complements to nouns.

A convenient way to analyze frequent language patterns in use has ensued because of the proliferation of data storage and analysis capabilities brought about by the technological revolution. The term *corpus* comes from the Latin root *corp-* meaning 'body,' and it has been commonly used in literary studies to refer to one author's body of work. In the present discussion, it refers to any electronically-stored collection of text. The Corpus under detailed analysis here is a unique compilation of single-authored, first-draft essays from freshman composition classes at The University of Georgia in the Spring semester of 2008 and shall be referred to as UGALECT.

According to Coxhead (2000), criteria for building a corpus include its representativeness (see also Biber, Conrad, & Reppen, 1998), organization, size, and the "criteria used for word selection" (Sinclair, 1991b, p. 215). A corpus is a 'principled' collection of texts, meaning the researcher(s) construct(s) the corpus with a particular

research agenda in mind, such as whether it is intended to be representative of speech, writing, or both. This study is focused on the more formal register of academic writing in NS freshman composition because of its informational application to the teaching of second language writing and academic vocabulary for L2 students at the college level.

As reported in the *Longman Grammar of Spoken and Written English* (Biber et al., 1999), the four structural categories most prevalent in academic writing are nouns and their cohorts: adjectives, determiners, and prepositions. Indeed, Halliday (1989) contends that lexical density in the form of elaborate noun phrases post-modified in various ways is especially characteristic of argumentative writing, which tends to report factual information. To date, most collocational studies have been done on the co-occurrence of content words such as nouns, adjectives, and verbs with each other even though prepositions are very high-frequency words many of which appear near the top of any frequency-derived list from a corpus of running text. Prepositions "make up about 8 percent of all the words we use in spoken English and about 12 percent of the words we use in written genres" (Kennedy, 2003, p. 246). In the British National Corpus (BNC), which includes 90 million words of written text "the most frequent 14 prepositions account for over 90 per cent of prepositional tokens in the corpus" (Kennedy, 2003, p. 247). ESL/EFL students would be well-served to have their attention drawn to the company these little, yet common, words keep (Sinclair, 1991b).

The empirical linguist, John Firth (1957), was the first to use the term 'collocation'[2] in corpus linguistics to refer to "lexical patterning along the syntagmatic

---

[2] For an account of the various historical uses of the term 'collocation' in linguistics, see Nesselhauf (2004b).

axis" (p. 196). Both Firth and Halliday (1991) advanced the notion that words have a statistical attraction to each other, i.e. a propensity for co-selection. In fact, a description of how words tend to co-occur was developed by Halliday and Hasan (1976) in their seminal work, *Cohesion in English*. Given one word, there is a 'calculable probability' that a certain other word will occur in its vicinity. A collocation is

> the way in which words are used together regularly…Collocation refers to the restrictions on how words can be used together, for example which prepositions are used with particular verbs, or which verbs and nouns are used together. (Richards, Platt, & Platt, 1992, p. 62)

Sinclair (1999) refers to such lexical choices as being either relatively open or restricted as determined by the grammar of the language. The more general term, 'phraseology,' has also been used to refer to the study of such "recurrent lexicogrammatical patterning" (Moon, 2007, p. 1045).

> *Lexicogrammatical* refers to frequently occurring combinations of words and grammar, where a particular word generally requires particular grammar. That is, the verb *required* can be followed either by an infinitive or by a *that*-clause. However, the most commonly used combination involves *required* followed by an infinitive. The combination of *required* and the infinitive is a *lexicogrammatical* pattern. (Coxhead & Byrd, 2007, p. 130fn).

The term 'colligation' is used to refer more specifically to the collocation of a particular lexical item with a particular grammatical word class such as a preposition.[3] In other words, the term 'collocation' refers to purely lexical relations, and the term 'colligation' refers to a relationship between lexical and grammatical words (Stubbs, 2001, pp. 64-65). Collocation frequencies in a corpus can be calculated by using a

---

[3] The term 'collocation' will be used henceforth when referring to two or more words frequently occurring together without regard to structure; the term 'colligation' will be used to refer to particular parts of speech frequently occurring together as in the case of N + P clusters.

concordancer software program, which locates and displays a targeted search term or phrase in its immediate lexical environments in a span of text, i.e. a concordance. The term KWIC is used to refer to a key-word-in-context, i.e. the node, which provides the axis or focal point in a list of concordances. In addition to displaying the actual lines of horizontal co-text, the advantage of displaying many concordance lines for a particular item simultaneously is the 'vertical dimension' (Sinclair, 2004), which can illuminate certain behavioral characteristics and regularities in the recurrences. For example, in Figure 1.1, the concordance window in AntConc 3.2.2w for *cost(s) of* when sorted alphabetically by immediate right and left collocates displays as:

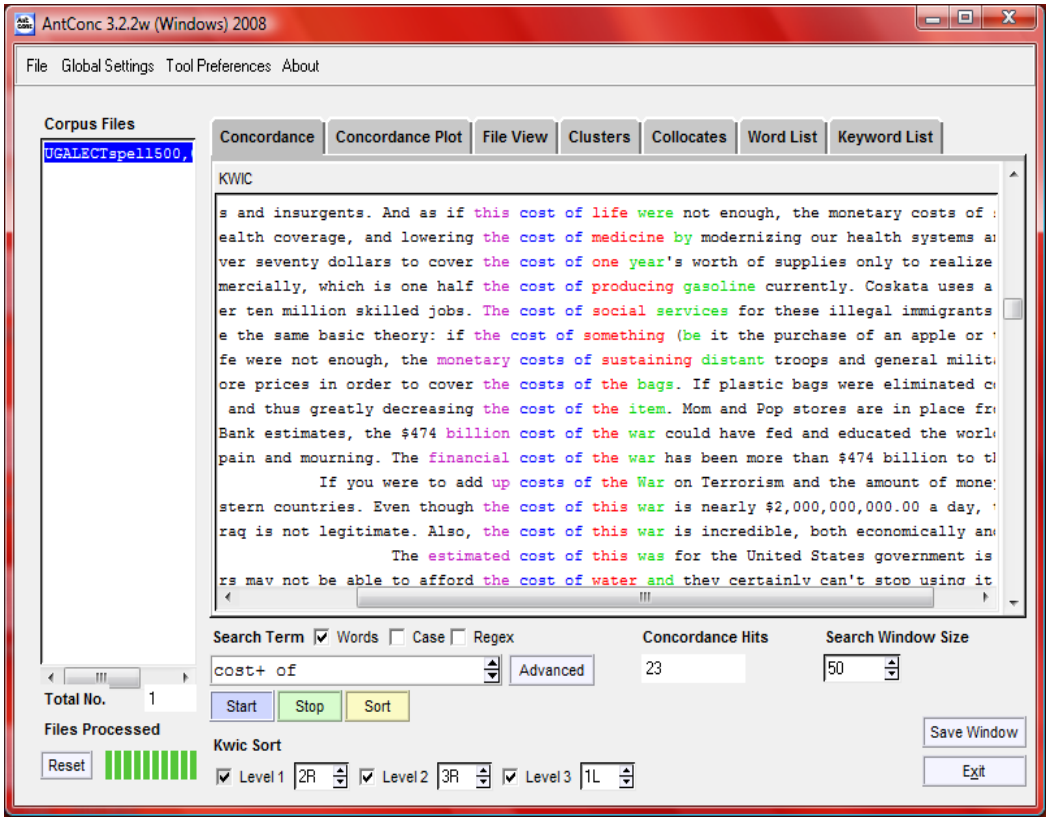

Figure 1.1 Concordances for *cost(s) of*

As can be seen clearly along the vertical (node) axis in this concordance list, *the cost of the/this war* is the most frequent contiguous collocation for the N + P cluster *cost(s) of* in the UGALECT Corpus. Indeed, a concordancer is a very powerful tool for discovering such recurrent patterns in actual language use.

'Formulaic language' is another term commonly used in the literature and refers to recurrent phrases having specific functions. Nattinger & DeCarrico (1992) define a lexical phrase as a 'pedagogically-applicable formulaic sequence.' In fact, they suggest giving lexical phrases a more central role in language pedagogy as a way to link the lexicon with the grammar of the language, as also promoted in Lewis' 'lexical approach' (1993, 1997, 2000) to teaching collocations. Such a lexico-grammatical focus offers a way to address both accuracy and fluency simultaneously by presenting learners with academic vocabulary in 'chunks' that have been validated by actual L1 usage (Pu, 2003).

A newer term, 'lexical bundle,' has been applied by Biber and Barbieri (2007) Biber and Conrad (1999), Biber, Conrad, and Cortes (2003, 2004), and Cortes (2002, 2004), among others, to refer to any frequently occurring string of three or more words in a sequence. These strings are discoverable using software that simply counts and lists in order of frequency all three-, four-, or five-word sequences in a given corpus without regard to grammatical structure. The resulting, statistically-frequent sequences are termed 'lexical bundles' and have been categorized by these researchers into a taxonomy by their various functions in discourse (see Biber, Conrad, & Cortes, 2003).

Some examples of lexical bundles from freshman composition are *is one of the, as well as the, at the end of,* and *to appeal to the* (Cortes, 2002). As can be seen here, lexical bundles often cut across two adjacent grammatical structures such as noun, verb,

and prepositional phrases. It is worth noting that a majority of the components of these lexical bundles consists of the less semantically-salient function words such as articles and prepositions, a mere function of their abundant utility in English. In fact, upon close scrutiny, lexical bundles seem hardly lexical. A more accurate description would be functional bundles, as the functional taxonomies outlined by their proponents indicate. Such frequent word sequences as characteristic of professional writers in academic disciplines have proven difficult to teach, even to native speakers in a writing-intensive history course (see Cortes, 2006). It is suggested here that a more targeted approach to discovering frequent patterns by formal categories in a corpus of written, academic discourse would serve to yield more fruitful information with regard to structural colligations, N + P, with the potential for applications in second language vocabulary and writing pedagogy.

To 'colligate' means literally to 'tie together,' and the term first appeared in the Oxford English Dictionary in 1953. 'Colligation' was later applied to corpus studies by Sinclair (1991b) and refers to the propensity for particular grammatical forms to co-occur, in this case nouns and prepositions. Prepositions in particular have rather stringent requirements with regard to their lexical environments. For instance, they must take a nominal complement whether explicit or implied such as in the case of so-called 'stranded' prepositions so frequent in spoken registers, especially in *wh-* questions: *Who(m) would you like the flowers sent to?* Furthermore, prepositional phrases as a whole always serve to modify other elements in a sentence, either as adjectival or adverbial complements. This characteristic makes them particularly suitable to a collocational analysis as detailed below.

Biber's 'complex noun phrases' are defined as nouns post-modified by prepositional phrases, relative clauses, participial phrases, and/or infinitive phrases. Prepositional phrases are the most common type of post-modifiers of nouns (Biber et al., 1999), and they can be thought of as a way to pack more information into short, dense phrases rather than with additional descriptive clauses. In fact, noun phrases with multiple post-modifiers are particularly characteristic of information-laden, formal language (Biber, 2006; Halliday, 1991; Reid & Byrd, 1998; Scott & Tribble, 2006). And the use of such a condensed code is indicative of the sophisticated, expository style that L2 writing students will want to emulate in their formal, academic discourse.

For the following analysis, an understanding of the notion of 'register' is essential. The term is used in linguistics to refer to a 'stylistic variety' of a language used in different situations with different persons and can be characteristic of different levels of formality. "A particular register often distinguishes itself from other registers by having a number of distinctive words, by using words or phrases in a particular way…, and sometimes by special grammatical constructions" (Richards, Platt, & Platt, 1992, pp. 312-313). C. C. Fries (1954) was one of the first to note that reading and speaking vocabularies are different and that word lists should be designed objectively based on frequency in the different realms of discourse, formal versus informal and written versus spoken. Large corpus studies have demonstrated that there is a quantifiable difference in the use of particular parts of speech and particular content words in different registers (Biber et al., 1999; Biber, Conrad, Reppen, Byrd, & Helt, 2002; Reppen, Fitzmaurice, & Biber, 2002). Because English prepositions are so common in general, and nouns are so much more common in formal academic registers, while at the same time nouns in

academic writing are frequently post-modified by prepositional phrases, this study hypothesizes that N + P clusters will have some significant role to play in freshman composition.

For Halliday (1991), a register consists of a set of probabilities of the occurrence of particular variables in a grammar. Those probabilities depend on the genre or text type, the register, the purpose, the environment, the situation, and so forth. So, no corpus can be truly representative of a language as a whole. But a corpus can be designed to represent some specific variety of language at some specific point in time in a particular place. This study analyzes the written expository and argumentative writings of native English speakers in freshman composition classes at The University of Georgia in the Spring semester of 2008. It is assumed that NNSs in freshman composition courses will want to write at least on a par with these native speakers.

1.5 Justification for the Study

As noted above, prepositions can make up to about 12% of written texts of English, and they are often misused in L2 writing, making for a distinctively non-native 'sound.' Articles and prepositions rank relatively high among frequency counts of ESL/EFL error types in second language writing studies (see Section 2.2). Articles and prepositions constitute "small but persistent problems" (Harris & Silva, 1993, p. 531) for even advanced, non-native speakers. Indeed, this seems to be a lingering struggle noted by L2 researchers themselves in their own writings in English (for example, see Li, 2008; Miller, 2007). And, basic L2 writers tend to "write in phrases patched upon phrases"

(Dykstra, 1997, p. 136) with little intra-sentential cohesion such as that which is provided by prepositions in their primary grammatical role as conjuncts of phrases.

Except for some very specific instances for designating spatial and temporal relationships, there are really few generalizable rules that can be resorted to in the use of prepositions. "Therefore, ESL writers need to learn prepositions the same way they learn other vocabulary items---through study or exposure to the language" (Harris & Silva, 1993, p. 535). Harris and Silva go on to recommend that ESL writing tutors, when addressing problems with grammar, should focus on verb tenses and inflections, inappropriate or missing prepositions, and missing articles as the most problematic areas for L2 writers. They further suggest that preposition problems are a result of 'limited lexical resources' about "knowing which one goes with a particular noun, verb, adjective, or adverb" (p. 534).

In an edited volume focusing on *Learner English* (Swan & Smith, 2001), each chapter presents the particular pronunciation, grammar, and vocabulary difficulties learners from particular L1 backgrounds may have based on contrasts between the L1 and the L2. For instance, there is a chapter on Spanish and Catalan speakers' common errors in English and a chapter on Korean speakers' common errors. Twenty-two chapters are each written by an expert on the L1 under analysis who is also a specialist in English language teaching. More than half of the chapters contain a separate section on difficult English prepositions for speakers of the respective native languages. A generalizable explanation is that because English has a relatively large number of prepositions compared to many other languages and makes finer distinctions in the distribution of their use, they are particularly challenging for virtually all second language learners.

When a single lexical item is equivalent to one or more lexical items in an L2, the difference is called a 'semantic split.' Semantic splits between the L1 and L2 were considered the most difficult for learners in the 'hierarchy of difficulty' outlined by Stockwell, Bowen, and Martin (1965) in their contrastive analysis of the grammars of English and Spanish. Basically, when learners have two or more choices in the L2, it was thought to be a much more difficult learning point than when the learner finds a semantic equivalency or merger in the L2. For example, a native English speaker will be likely to have some difficulty early on in discerning the various uses of *por* and *para* in Spanish because they are generally equivalent to one word in English, *for*.[4] According to contrastive analysis, native English speakers would have less difficulty with *en* in Spanish because it represents a 'semantic merger' of two English prepositions, *in* and *on*. This oversimplified view of L2 learning has been largely discredited as the picture turns out to be much more complicated than first realized. Sometimes a greater degree of difference from the L1 can actually facilitate learning as this difference makes the point more salient to the learner. Often, it is those cases of more subtle distinctions between the conventions of two languages that prove most challenging to learners. Prepositions are subtle. Hence, the distribution of use for the two prepositions *in* and *on* in English could be addressed by learning them in their greater contexts, as collocates to other, more salient content words in the common patterns of their respective L1 uses. Indeed, this is the way Azar (2003), a very popular ESL/EFL grammar textbook series, presents certain preposition combinations with adjectives and verbs such as *capable of* and *believe in*.

---

[4] For a longitudinal examination of L2 acquisition of the Spanish prepositions *por* and *para* by L1 English speakers, see Lafford and Ryan (1995).

Particularly confusing for native Spanish speakers are the preposition distinctions in English among *in/on/into, to/at/in, as/like, for/by,* and *during/for* (Coe, 2001, pp. 108-109), some of the most common words in the language. A semantic approach to prepositions simply fails to clear the air because they can be highly idiomatic, and sending a student to a conventional dictionary may even exacerbate the problem because definitions for prepositions tend to be some of the longest due to their wide distribution of use. A collocational approach, on the other hand, serves to draw attention to the most common environments for each high-frequency preposition. In many cases and with many native languages, there is simply no one-to-one correspondence with English prepositions, and collocations represent patterns in the target language that serve to characterize particular registers.

As described in Thornbury (1999), approaches to grammar can be deductive, with a focus on general, abstract rules which are then filled in with concrete vocabulary items in a piecemeal fashion, or inductive, with a focus on specific examples from which researchers, materials writers, teachers, and even students can embark on a process of discovery, uncovering the patterns of the L1 as currently used by NSs. In cautioning against any extreme methods, Widdowson (1989) concludes that

> the structural approach accounts for one aspect of competence by concentrating on analysis but does so at the expense of access, whereas the communicative approach concentrates on access to the relative neglect of analysis. (p. 132)

The communicative approach to language teaching has been very popular for several decades now, but it remains inefficient in that it takes little advantage of the patterns and conventions inherent in written academic language, and it downplays the useful,

analytical abilities that adult learners bring to the task. Howarth (1996) points to the fact that little focus has been placed on form:

> In recent years the dominance of the communicative approach in the teaching of English as a foreign language has tended to place much greater emphasis on learners' ability to use their vocabulary resources creatively in order to 'negotiate meaning' spontaneously, and this approach has consequently had little interest in studying prefabricated language. (p. 134)

Furthermore, communicative language teaching methods tend to focus on the oral language, which is demonstrably different from formal, written language conventions. In deference to a greater focus on the transfer of meaning, communicative methods have also ignored the significance of the most frequent, small words of English. In his lifelong dedication to corpus-based linguistic research for the benefit of L2 teaching, Sinclair (1991a, 1991b, 1999, 2004) championed the importance of small words because of their ubiquitous nature in English. In that same vein, *The Lexical Approach* (Lewis, 1993) views lexis as primary and interdependent with grammar in its focus on teaching collocations, especially collocations with *of*, which has been shown to play a central role in the post-modification of noun phrases (see Appendix C). "In many examples…*of* is closely related to the word which **precedes** it rather than the word that follows it, so at best the term 'preposition' is highly inappropriate. Nor is it [*of*] typically about possession" (Lewis, 2000, p. 145). The term itself, pre-position, indicates the close ties these words have to their following objects and downplays their intrinsic relationship to any words they actually modify, their predecessors. In fact, Scott and Tribble (2006) found the form N + *of* to occur in over 79% of instances of *of* in the written academic portion of the British National Corpus, whereas this pattern occurred in just less than

50% of such instances in conversational productions in the same corpus. *Of* itself here presents a

> significant contrast between Written Academic and Conversational Production – the immediate left collocates of *of* in Conversational Production constitute a very small set of words with the top five *sort, bit, one, lot,* and *out* making up 40% of the total, and the top 20 accounting for 71% of the total instances…Even where there are instances of post-modifying *of* in Conversational Production, it tends to be in the context of fixed, highly generalised phrases, and spans an extremely small set. In extreme contrast, in Written Academic the top 20 left collocates of *of* constitute a much smaller percentage of the total instances (23% - with the top five only representing 10%). From a language teaching perspective, this set of collocates of *of* in Written Academic is also significant in that it offers at least two potentially useful insights for learners. The first is that it provides a starting point for a review of the prefabs that were used by this set of writers (and which are likely to be important for other academic writers). *terms of, range of, form of, case of, principle of, effect of, function of* are all potentially valuable to apprentice writers. Secondly, it could be used as the starting point for even narrower disciplinary investigations of the left collocates of *of*. (p. 100)

Collocations can provide direct access to the present-day conventions of preposition usage in English by presenting them as components of larger lexical units. The vast amount of quantitative data with regard to linguistic patterns that can be garnered from NS corpora remains an under-exploited resource for informing second language pedagogy. In what follows, it is argued that some N + P clusters should take their rightful place alongside multi-word verbs and prepositional adjectives as viable and robust lexical units warranting consideration in ESL/EFL textbooks and classrooms.

1.6 Purposes for the Study

The specific purposes for the present study are:

- To review the field of L2 academic vocabulary
- To review the field of corpus-based analyses of academic writing
- To present evidence of ESL/EFL errors with regard to prepositions in SLW
- To discover the most frequent N + P clusters in NS academic essays
- To sanction the consideration of robust N + P clusters as viable lexical units

1.7 Research Questions

In a qualitative analysis of learner errors with prepositions, the researcher asks: What types of errors do L2 learners make with regard to English prepositions in their academic writing?

From the 500,000-word Corpus of first-draft, native speaker, freshman essays (UGALECT), the following quantitative research questions will be addressed:

- What are the most frequent prepositions used by native speakers in freshman composition?
- What are the most frequent nominal left colligates of the ten most frequent prepositions in freshman composition, and what are the frequencies of occurrence of these two-word phrases (N + P clusters) in the UGALECT Corpus?
- Are these nouns usually followed by prepositions in the Corpus, and, if so, which prepositions are their most frequent right colligates? In other words, what proportion of these nouns is post-modified by a particular preposition as opposed to some other preposition?
- Do other frequent prepositions prove to be as useful as *of* as nominal right colligates in the written academic register of native speakers?

- Do the most robust N + P clusters in the NS essays occur in the NNS essays?
- What is the nominal density of the NS academic writing compared to the NNS academic writing? Does the learner data exhibit the same nominal density as the native speaker data? What about preposition density?

In this descriptive analysis, the researcher seeks to highlight robust N + P clusters in written academic English that may warrant some attention in L2 academic vocabulary presentations. In the spirit of Coxhead's Academic Word List (2000), the researcher hopes to sanction an academic phrase list[5] for use by L2 materials writers, teachers and students.

[5] For a statistical analysis of two-word clusters based on the Academic Word List, see Coxhead & Byrd (forthcoming) from Michigan University Press, *The AWL: Collocations and recurrent phrases*.

CHAPTER 2

REVIEW OF THE LITERATURE

This chapter presents the relevant literature on L2 academic vocabulary, L2 preposition errors in academic writing, and corpus findings in variation studies with regard to collocations and 'lexical bundles' involving nouns and prepositions in L1 academic writing.

2.1 L2 Academic Vocabulary: English Word Lists

There is a long tradition of generating academic word lists for educational purposes based on frequencies in academic discourse (Campion & Elley, 1971; Coxhead, 1998, 2000, 2002; Fries & Praninskas, 1972; Thorndike, 1932; Thorndike & Lorge, 1944; Traver, 1950; West, 1953; Xue & Nation, 1984). Thorndike (1932) first provided a list of 20,000 common content words for teachers of English, which was later expanded to 30,000 words (Thorndike & Lorge, 1944). Academic vocabulary teaching has usually focused on content words because they carry the greatest semantic weight. Such lists consist of nouns, verbs, and adjectives with high frequencies in English, and frequency and range, or distribution of use, have long been thought of as a way to rank words by their relative significance for English language learners. In fact, the two-thousand most frequent words in a 10-million word corpus of written and spoken English were found to

account for 83% of the entire text (O'Keeffe, McCarthy, & Carter, 2007), so students would be well-advised to focus on these common words first in a lexical syllabus.

The General Service List (GSL) consists of about 2000 'headwords' (West, 1953), which are stem noun or verb forms. Because it was based partly on raw frequencies in a five million-word corpus, the GSL did include function words such as articles, prepositions, conjunctions, and pronouns, most of which can be found near the top of the list. The GSL also considered semantic relationships among various forms and organized content words around headwords for the purpose of alerting students to the many inflected forms a word can take in a sentence. Nation (1990) includes a list of content words from the GSL not likely to be well-known by pre-university ESL students based on translation tests. Words not known by any of the students tested include the common nouns *account, approval, course,* and the prepositional phrase *in spite of*. Also, Xue and Nation (1984) presents a University Word List (UWL), which contains the following frequent and widely distributed nouns: *alternative, component, region, role, status, summary, technique,* and *usage* (pp. 235-239). Each of these nouns could reasonably be followed by at least one of the top-ten prepositions of English: *alternative to, component of, role in,* and so forth.

Coxhead proposed the Academic Word List (AWL) as a "useful example of corpus-based research leading directly to teaching and learning applications" (2002, p. 79). With this list, Coxhead hoped to replace the UWL (Xue & Nation, 1984) because she felt the earlier list was based on too small and varied a corpus, and Coxhead specifically wanted to go beyond the first two-thousand words in West's GSL (1953) by composing her list from a 3.5-million-word corpus containing academic writing from

four different disciplines: arts, commerce, law, and science. Coxhead (2002) contends that the AWL consists of the most relevant, useful, and frequent content vocabulary for students pursuing higher education in an English-speaking environment, and several textbooks on ESL/EFL vocabulary have ensued with a focus on contextualizing the 570 'word families' on this list.[6] In justifying the need for an academic word list, Coxhead (2000) believes that "academic words…are not highly salient in academic texts, as they are supportive of but not central to the topics of the texts in which they occur" (p. 214). Therefore, by way of word lists the attention of language students can be explicitly drawn to words they may have paid little attention to in their academic reading. Simple word frequencies in a large sampling of particular text types, in this case academic writing, can reveal to us just these types of wide-ranging, non-topical vocabulary items specific to academic and more formal registers. Coxhead (2000) tested her AWL for occurrences in fiction and found a very low correlation (1.4%) with these academic content words, further establishing the need for, and status of, these items in higher education, where a great deal of non-fiction writing will be encountered by students.

Schmitt (1997, 2000, 2004) is largely responsible for making these vocabulary lists more accessible for teaching and learning purposes in applied linguistics[7] and has developed tests based on the AWL, which can serve to place learners in appropriate academic levels. In his discussion of collocation, Schmitt (2000) notes that "vocabulary choice is constrained by systematicity" (p. 76). Not only must words co-occur to be considered collocates, but there must also be some degree of exclusivity. For example,

---

[6] For example, see the *Academic Word Power* series from Thomson Heinle.

[7] See Schmitt and Schmitt (2005) for an ESL textbook based on the AWL.

he observes that the article *the* can co-occur with almost any common noun, so this would not be considered a collocation (p. 77). On the other hand, the notion that preposition choice may be determined by an immediately preceding noun is a principle that has yet to be exploited in L2 teaching. "Grammatical collocations are the type in which a dominant word 'fits together' with a grammatical word, typically a noun, verb, or adjective followed by a preposition. Examples are *abide by, access to,* and *acquainted with*" (Schmitt, 2000, p. 77). Schmitt (2000) regards collocational investigations as one of the most important new directions in vocabulary studies with "the realization that words act less as individual units and more as part of lexical phrases in interconnected discourse…[and] lexical phrases in language reflect the way the mind tends to 'chunk' language in order to make it easier to process" (p. 78). Further, if such items are stored as lexical units, should we not also teach them as such?

Nation, the foremost authority on second language vocabulary, contends that "many linguists now consider the lexicon to play an important, if not central, role in grammar" (2001, p. 55). He agrees with Sinclair (1991b) in that part of knowing a word is knowing which other words it may be used with, and that by teaching such word patterns, the learning burden can be reduced for certain words. Academic vocabulary lists are considered significant because they account for not only a large number of these words, but also for the vocabulary in a wide range of academic texts (Nation, 2001, p. 189). However, such word lists are in need of contextualization, and phrase lists are a

step in the right direction.[8] Nation (2001, p. 319) offers the following examples of mental chunking at different linguistic levels for written language:

Table 2.1 Types of Chunking

LEVEL	**Type of Chunking for *play***
Letters	The letter *p* is processed as a unit, not as a set of 2 separate strokes.
Morphemes	The morpheme *play* is processed as a unit, not as a set of 4 letters.
Words	The word *player* is processed as a unit, not as a set of 2 morphemes.
Collocations	The collocation *player with promise* is processed as a unit.

The notion of mental chunking remains to be proven valid as a psycholinguistic reality,[9] but the notion of presenting learners with more efficient ways to master both the lexicon and grammar through frequent collocations of English is a promising direction for corpus linguistics studies. Prepositions are a significant word class in English simply because they are so prevalent as linking devices, but academic word lists as noted above fail to include any consideration of these abundant little words. Except for their presence in multi-word verbs and in transitional prepositional phrases such as *of course, in fact,* and *on the other hand*, they receive little attention in second language vocabulary and writing instruction. With regard to utility,

> grammatical words are necessary to the structure of English [sentences] regardless of the topic, …[and] one of the reasons L2 learners do not sound native may be that they overuse certain relatively infrequent words and underuse certain relatively frequent words. (Schmitt, 2004, p. 73-76)

Even advanced, second language writers have distinct difficulties with using and selecting appropriate prepositions as evidenced by the many studies that have been done

---

[8] In fact, Coxhead & Byrd are currently working on just such an analysis of two-word clusters based on the Academic Word List (Byrd, personal communication).

[9] See Sosa & MacFarlane (2002) for an examination of the holistic storage of and access to two-word collocations involving the word *of* following the usage-based model of the lexicon (Bybee, 2001 & 2002).

on error frequencies in second language writing (Benson et al., 1992; Ene, 2007; Flowerdew, 2006; Hemchua & Schmitt, 2006; Jiménez-Catalán, 1996; Khampang, 1974; Meziani, 1984; Neff, Ballesteros, Dafouz, Martinez, & Rica, 2004; Reid, 1988).

2.2 Errors in Advanced L2 Writing

Reid's (1988) doctoral dissertation was an early quantitative corpus study contrasting the use of particular linguistic structures in the academic prose of native speakers of English with that of various non-native speakers, including students from Chinese, Spanish, and Arabic L1 backgrounds. Table 2.2 is a generalized representation of Reid's statistically significant findings with regard to the use of "selected cohesion variables" (p. 82) such as pronouns, conjunctions, and prepositions:

Table 2.2 Relative Usage of Function Words in L1 & L2 Writing

VARIABLE	ENGLISH	SPANISH	ARABIC	CHINESE
Pronouns	Low	High	High	High
Conjunctions	Low	High	High	High
Prepositions	High	Low	Low	Low

What is interesting here are the quantitative differences in the use of function words between the native and all of the non-native speakers. The native speakers used a relatively low percentage of conjunctions and pronouns in comparison to all of the non-native speakers, and the native speakers used a relatively high percentage of prepositions in comparison to all of the non-native speakers. This finding indicates that non-native speakers who are being taught to write academic English may need some specific direction in the area of preposition usage as appropriate to such informationally-dense writing. Also, the learners' relatively high usage of pronouns could indicate a vocabulary

deficiency with regard to nouns. Reid goes on to say that Biber (1985, 1986) found that formal, informational writing is marked by a limited use of pronouns for native speakers, yet non-native speakers tend to overuse them, possibly because of a lack of content vocabulary (Reid,1988). Biber (1988) also contends that formal, informational writing is characterized by a preponderance of complex noun phrases, which are those followed by multiple post-modifiers such as prepositional phrases. This observation suggests that some attention to this deficit in non-native speaker academic writing is warranted.

Reid (1988) proposed that a greater reliance on pronouns might indicate a lack of nominal vocabulary on the part of learners. And, the fact that there are only seven coordinating conjunctions in English may render this class of items relatively easy to master for second language students, *and, but,* and *so* being by far the most frequent and semantically transparent. Prepositions, on the other hand, come in a variety of forms with varying degrees of semantic opacity. Reid's study demonstrates that learners of several, vastly different L1s do not utilize English prepositions in their academic writing to the same extent as native speakers do even at advanced levels.

In her examination of the academic writing of eleven, non-native graduate students in applied linguistics, Ene (2006) found that they made the most writing errors with regard to articles followed at some distance by prepositions and then nouns (p. 398). These are all word classes associated with written language:

- articles *a* and *the*, indicating a high instance of noun phrases
- the preposition *of*, suggesting post-modified noun phrases…
- prepositions *to, for,* and *in*, suggesting prepositional phrases. (O'Keeffe, McCarthy, & Carter, 2007, p. 12)

Function words were a particular weakness in Ene's advanced learners' writing even though they were studying to be English language teachers themselves.

In another study contrasting native and non-native writers, Benson et al. (1992) found that Basic (NS) Writers at the college level did not make the same kinds of grammatical mistakes that second language writers made. The Basic Writers averaged fewer errors specifically with regard to verb tenses, articles, and prepositions. This finding suggests that errors with these particular forms can be indicative of non-native speaker usage. In fact, Henning (1978) felt that difficulties with "standard prepositions" in the college writing of Iranian students may be indicative of their level of mastery of L2 English (p. 387). Bitchner, Young, and Cameron (2005) found that although corrective feedback was successful at improving accuracy with regard to writing errors such as the simple past tense and the definite article, prepositions remained problematic for their learners. Even when preposition errors are marked as such, students have difficulty correcting them without specific corrections provided. Also, with regard to feedback, Lee (2004) emphasizes that students are reliant on writing teachers for comprehensive feedback. If comprehensive feedback is not provided, students will assume their usage is accurate.

In a study of lexical errors in the academic writing of Thai learners, Hemchua and Schmitt (2006) developed a comprehensive error taxonomy. Second only to 'near synonym' errors, i.e. word choice, which is also a collocational issue, prepositions and suffixes were found to cause the greatest degree of difficulty (p. 3). These researchers consider the sources of these errors as more due to the 'intrinsic difficulty' of the L2 English rather than to any L1 transfer.

In a study of error gravity in Israeli EFL student writing, Salem (2007) supports the notion of the interdependency of grammar and lexis. Although lexical errors were deemed more serious than grammatical errors, the interplay of grammatical accuracy with lexical choice is evidenced. This study highlights the fact that certain content words entail certain grammatical words in English colligations, and without this kind of phrase level knowledge, students may choose awkward, or even omit, appropriate prepositions.

With regard to native Spanish speakers' academic writing in English, Neff et al. (2004) found most lexical errors (23%) involved prepositions or adverbs. Many of the error examples demonstrate collocational problems, which the authors attribute to a "lack of reading in English, a major source of input for collocations" (p. 216). Their students had particular difficulty with confusion between *in* and *on,* which coincide with one word in Spanish, *en.* Germany and Cartes (1995) demonstrated that most errors in the EFL writing of Chilean students that they analyzed with regard to English prepositions of location were due to L1 transfer and the abstract qualities of certain English prepositions, especially *at, in,* and *on* (p. 44).

Jiménez-Catalán (1996) also points out the high rate of errors with English prepositions for native Spanish speakers. She contends that English language textbooks fail to emphasize that "a given preposition has more than one meaning depending on the context or that some verbs require an obligatory preposition" (p. 172). In 290 essays written by secondary school students, this study found substitution by a different preposition, such as in *There was a lot of money into the handbag*, to be the most frequent error type, at about 12%, followed by noun and verb substitutions. Also, addition and omission of prepositions occurred in another 7% of the error types.

Preposition substitution errors were made by 75% of the students, and addition/omission errors were made by one-third, with *of* being the most frequently appended and *to* the most frequently omitted preposition. Jiménez-Catalán (1996) contends that such problems with English prepositions are not restricted to native speakers of Spanish nor "to any particular group of students since the foremost position of preposition errors in lists of the most frequent error types compiled from learners of English of different nationalities has been reported by researchers in the field" (p. 171). In fact, in a diagnostic test on the prepositions *at, by, for, from, in, on, to,* and *of* administered to Thai, Japanese, and Spanish-speaking students, Khampong (1974) found no significant differences in the groups' scores (p. 215). In other words, no items could be distinguished as specifically Thai problems with English prepositions. Also, in looking at English speakers' L2 Spanish, Azevedo (1980) showed that choice of preposition remains 'imperfectly mastered' by graduate students who were at an advanced level of Spanish.

In a learner corpus analysis of Chinese students' academic writing in English, Flowerdew (2006) found the most frequent error type (68%) that learners made with regard to 'signalling nouns,' which he defines as those nouns "which have cohesive properties across and within clauses" (p. 345), was in their colligations with following prepositions. He provides the following examples of the Chinese students' misuses of English prepositions following nouns: **argument in* rather than *argument for*, **chance to* (inf.) rather than *chance of*, **discrimination to* rather than *discrimination against*, **effort on* rather than *effort to* (inf.), **argument on* rather than *argument for*.

In a comparison of the error corrections made by EFL writing instructors who were native speakers of English and those who were native Japanese speakers, the latter

group was found to have overlooked errors involving articles, prepositions, and loanwords from English (Kobayashi, 1992). Thus, even advanced non-native speakers who teach EFL may continue to have difficulty recognizing errors with English articles and prepositions.

2.3 Corpus Studies of NS English Usage

George Kingsley Zipf was a Harvard professor of psychology during the middle of the 20th century who was interested in certain manifestations of speech, especially that of children (Zipf, 1942) and schizophrenics (Whitehorn & Zipf, 1943). Through corpora analysis, Zipf (1945a) was able to come up with a mathematical formulation regarding the rank/frequency relationship of words in running text:

> As far as the general frequency of occurrence of words is concerned, it has perhaps always been known by students of speech that a few words occur frequently while many (indeed most) occur rarely---a relationship that has become ever more striking as a result of the accumulation of detailed frequency lists of words for many languages as compiled by students of spelling, stenography, linguistics, and psychology. (p. 127)

According to what later became known as Zipf's Law, the frequency of any word in a corpus of naturally-occurring text is inversely proportional to its rank in that frequency (Zipf, 1945b). In other words, an item's rank order in a frequency list multiplied by that item's actual number of occurrences tends to remain constant. For example, the most frequently occurring word, which is usually *the* in English, occurs about twice as often as the second most frequent word, which occurs approximately twice as often as the third most frequent word and so on. In the Brown Corpus (Kučera & Francis, 1967) of one million words of American English, *the* makes up almost 7% of the

text, and *of*, the second most frequent word, comprises just over 3.5%. In fact, "only 135 vocabulary items are needed to account for half the Brown Corpus" (http://en.wikipedia.org/wiki/Brown_Corpus).

A number of corpus studies have been done especially over the last decade (and especially in Europe) for the primary purpose of informing second language pedagogy. John Sinclair has been described as the father of corpus linguistics. He was primarily responsible for the Cobuild Project of the 1980s, which resulted in an exhaustive, corpus-based, multi-volume dictionary for English language learners. The basic premise of his work is that the most frequent linguistic behavior of native speakers would be very useful insight for learners of the language, and he promoted a move towards data-driven learning (DDL, see also Johns, 1994; Scott & Tribble, 2006), whereby students are instructed in tasks designed to utilize the resources of corpus linguistics in conjunction with the now readily-available amount of data in the form of electronic texts on the internet as a way to discover for themselves how present-day English really works.

Sinclair (1991b) was one of the first to recognize that a large percentage of the language we use consists of 'prefabricated chunks.' Such chunks reside along a collocational continuum of relatively fixed and relatively free word combinations in the language. He proposed the 'idiom principle' at one end of the continuum to account for most language production, in which lexical choices are restricted by the language, and the 'open choice principle' at the other end to account for unique word combinations and

idiosyncratic usage (Flowerdew & Li, 2007[10]). The idiom principle asserts that phrases or "strings that would appear to be analyzable into segments nevertheless constitute single choices" (Erman, 2007, p. 25) for the language user. In support of Sinclair's idiom principle, Erman and Warren (2000) contend that both spoken and written language is made up of a large amount of these prefabricated chunks.

Sinclair (1991b) asserts that traditional grammars tend to be guided by the open choice principle, whereas most actual language usage is quite restricted by the lexico-grammar of the language. (p.110). According to Howarth (1998), at the open end of the continuum, we have free combinations such as *under the table*, in which lexical choice is quite variable; at the 'pure idiom' (or fixed) end of the continuum we have *under the weather*, which has "a unitary meaning that cannot be derived from the meaning of the components" (p. 28). Along the middle of the continuum, we have *under the microscope*, which is a 'figurative idiom,' i.e. a metaphor, and somewhat restricted, and we have *under attack* as a more 'restricted collocation' (Howarth, 1998). Sinclair maintains that it is these forms along the middle of the grammar continuum that cause the most difficulty for students because free combinations at one end are unrestricted and true idioms at the other are relatively rare (also noted by Biber et al., 1999).

> If it is the case that the node word occurs with a span of particular words at a frequency greater than chance would predict, then the result is a collocation. The more certain the words in a span are to co-occur, the more fixed and idiomatic the collocation. With completely fixed

---

[10] Flowerdew and Li also point out here that Sinclair's idiom principle is what antiplagiarism devices are based on. The probability for the recurrence of any word sequence is exponentially decreased by the length of that sequence. For example, four-word sequences are ten times more likely than five-word sequences (Biber et al., 1999). The longer the sequence, the less likely it is to be repeated. Therefore, the repetition of any four- word sequence or above in a corpus of running text is highly unlikely. Contiguous collocational recurrences of any length are significant, i.e. lexical bundles..

> collocations such as many idioms and clichés, mutual expectancy has become fixed, syntagmatically and paradigmatically ossified, which results in loss of meaning because of elimination of an element of choice. As collocations become less fixed, that is, as more variation becomes possible along both axes, predictability lessens and meaning increases. (Nattinger & DeCarrico, 1992, p.20)

Prepositions play a large part in contiguous collocational sequences because of their essential role as connectives among phrases in a sentence.

Kennedy (2003) acknowledges the difficulty of prepositions for non-native speakers in his guide to the structure and meaning of English for second language teachers:

> Prepositions are by common consent one of the hardest parts of English to learn how to use. There are about 100 prepositions. They make up about eight per cent of all the words we use in spoken English and about 12 per cent of the words we use in written genres…Research on large corpora has shown that a small number of prepositions account for most occurrences. (pp. 246-7)

Kennedy also provides a list of the distribution of prepositions in the written portion of the British National Corpus (BNC), which was composed of over 90 million words at the time. The top fifteen along with their relative percentages are:

of	26.1	*with*	5.7	*as*	1.9
in	16.1	*by*	4.6	*into*	1.4
to	8.1	*at*	4.1	*about*	1.1
for	7.3	*like*	3.8	*after*	1.0
on	5.7	*from*	3.7	*between*	.8

Thus, *of* makes up more than a quarter of all the prepositions in this extremely large corpus, and just the top three prepositions account for half. Because corpus research has shown that a small number of prepositions can account for most occurrences of prepositions, this study will focus on only the ten most frequent prepositions in the NS Corpus under analysis.

2.4 Lexical Bundles in Academic Discourse

The following studies on lexical bundles, in which many N + P clusters occur, serve to inform the present study with regard to the quantitative differences in spoken and written registers.

In a corpus comparison of the frequency of word classes and functions in use across various university registers, Biber et al. (1999) found that nouns and their colligates, which consist of determiners, adjectives, and prepositions, are more common in news reports and academic prose and less common in conversation, where more verbs and adverbs abound. Biber (1988) describes prepositions in particular

> as an important device for packing high amounts of information into academic nominal discourse…Prepositions tend to co-occur frequently with nominalizations and passives in academic prose, official documents, professional letters, and other informational types of written discourse. (p. 237)

In fact, prepositions frequently co-occur with nouns in written, informational discourse in general (Biber, 1988). Biber's studies have focused on what he calls 'lexical bundles' (introduced in Section 1.4 above), which can be defined as three or more words occurring frequently together in a linear sequence. Lexical bundles can be thought of as contiguous collocations because they involve a sequence of words. A computer software program simply records each and every occurrence of a word and the two (or more) words following it in a corpus and counts the frequency of each such bundle to come up with the most common. In order to be included in the results as a lexical bundle, the series has to occur at least 20 times in one-million words and in five or more different texts in order to exclude possible idiosyncratic uses by any individual author (Biber, 1988).

Using a representative corpus of text in a university setting of 5 million words per register, Biber et al. (1999) provides an extensive quantitative and contrastive analysis of the use of particular parts of speech in each register. The following chart is a binary depiction based on Biber et al. (1999) of the relative prevalence of certain parts of speech in the different registers of speech and writing, all as used in a university environment:

Table 2.3 Relative Frequencies of Word Classes in Academic Discourse

	CONVERSATION	FICTION	NEWS REPORTS	ACADEMIC PROSE
More Common	Pronouns Verbs/Adverbs Auxiliaries Particles	Pronouns Verbs/Adverbs Auxiliaries Particles	Nouns Adjectives Determiners Prepositions	Nouns Adjectives Determiners Prepositions
Less Common	Nouns Adjectives Determiners Prepositions	Nouns Adjectives Determiners Prepositions	Pronouns Verbs/Adverbs Auxiliaries Particles	Pronouns Verbs/Adverbs Auxiliaries Particles

More specifically, Biber et al. (1999, p. 996) found that 4-word lexical bundles realized as a personal pronoun plus a lexical verb phrase, such as *I don't know what…*, made up 44% of four-word lexical bundles in the conversation register and did not factor in the written registers at all. In the written academic register, however, 30% of 4-word lexical bundles consisted of a post-modified noun phrase such as *the nature of the…*, and 33% of 4-word lexical bundles consisted of a preposition plus a noun phrase fragment such as *as a result of….* This abundance of nouns and prepositions in the written, academic register motivates the focus of the present study on N + P clusters for second language writers.

Also, Biber et al. (1999) finds a reciprocal relationship between the use of certain function words and certain content words:

> The distribution of function words is closely connected with the distribution of lexical word classes…The low frequency of nouns in conversation is compensated for by the high pronoun density. Conversely,

> a high frequency of nouns in news and academic prose corresponds to a low density of pronouns…Conversation and fiction have the highest frequency of lexical verbs and also the highest frequency of auxiliaries and adverbial particles, which specify or extend lexical verbs. Similarly, function words associated with nouns vary in frequency with the density of nouns. Academic prose and news reportage have the highest frequency [of nouns]. (Biber et al., 1999, p. 92-93)

Thus, we can reasonably suppose that students at U.S. universities will be exposed to the kind of nominally-rich language expected of them in formal, academic writing only insofar as they read academic prose (textbooks) and/or news articles. Otherwise, just as with native speakers, their writing could be marked by features of the conversational register (such as pronoun density) to which they are exposed.

In her criticism of extant ESL grammar curriculum guidelines, Byrd (1998) was also able to make a number of similar observations with regard to part-of-speech frequencies based on a corpus analysis of academic textbooks. Such writing is inherently designed to convey large amounts of "information including data, theory, definitions, and other types of generalizations about habitual behaviors and the natural world" (p. 91). As for the use of particular grammatical structures in this type of information-laden writing, she shows that it is characterized by the use of (in order of relative frequency):

> -long, complicated noun phrases
> -generic noun phrases…to refer to categories rather than to individuals
> -passive verbs
> -a limited set of verbs
> -present tense (to discuss habitual behavior, scientific facts, or general truths).

Byrd feels it would serve our students well in the second language writing classroom to focus the grammar curriculum on just such structures. She goes on more specifically about the structure of complex noun phrases in academic prose in particular:

> Long, complicated noun phrases are often used as is specialized terminology. The complexity of the noun phrase involves 1) strings of

> adjectives and nouns in front of the core noun, 2) relative clauses attached to the noun and often reduced to participle phrases, and/or 3) strings of **prepositional phrases after the noun**…Because the emphasis [in such writing] is on theory, facts, and concepts rather than on human beings, *it* is the most commonly used personal pronoun. On the other hand, this type of material often repeats the same noun phrase rather than using a pronoun to refer to it --- possibly because of the importance of using exactly the correct terminology. [In contrast]…the range of lexical verbs and of verb tenses is narrow in comparison with conversational or narrative uses of English. (Byrd, 1998, p. 91) [boldface added]

Both Byrd (1998) and Biber et al. (1999) highlight that a distinctive property of the written academic register is a preponderance of complex noun phrases and the post-modification of those noun phrases in the form of prepositional phrases. "In academic prose, over 60% of all lexical bundles are parts of noun phrases and prepositional phrases" (Biber et al., 1999, p. 995).

Cortes has also focused her corpus studies on lexical bundles in academic writing, both in freshman compositions (2002) and in history and biology textbooks (2004). Cortes' list of 4-word lexical bundles found in NS freshman writing is provided below:

a lot of the	*at the same time*	*the back of the*
a part of the	*in an effort to*	*the bottom of the*
a wide range of	*in the case of*	*the edge of the*
a wide variety of	*in the form of*	*the side of the*
as a result of	*in the United States*	*to appeal to the*
as well as the	*is one of the*	*to be able to*
at the bottom of	*it is as if*	*will be able to*
at the end of	*it is difficult to*	
at the top of	*on the other hand*	

Topic specific bundles and those representing titles of narratives being analyzed in the composition classes were excluded from the list. As can be seen, lexical bundles do not represent any 'complete structural units,' and Cortes notes that Biber and Conrad (1999) found that "less than 5 percent of lexical bundles identified in academic prose can be regarded as complete grammatical units" (Cortes, 2002, p. 135). Thus, rather than

designating them by structure, Cortes categorizes lexical bundles with regard to function, setting up a taxonomy of their usage as organizers of discourse. She also found no one-to-one correspondence between lexical bundles as expressed in L1 English and in L1 Spanish even though they may have the same function (Cortes, personal communication). In other words, both writers in Spanish and writers in English find similar rhetorical reasons for utilizing frequent lexical sequences though, as we might expect, those sequences vary in structure or form even when expressing the same meaning.

What is immediately apparent from this list is the preponderance of nouns and their colligates: articles and prepositions. Indeed, Cortes calculates that 35% of these lexical bundles found in freshman writing are noun phrases with a post-modifier fragment (almost all of which are prepositions), and 30% are prepositions plus a noun phrase fragment. This means that well over half of the bundles involve some segment of prepositional phrases. This fact, along with Biber's findings that post-modified noun phrases are especially dense in academic writing, also motivated this study on prepositions and their nominal left colligates, as they may be considered especially relevant structures for non-native speakers learning to write at the college level.

The most common lexical bundle in Cortes' data by far was *in the United States*[11] with 141 occurrences in 306,704 words. With regard to grammatical group function, Cortes divides prepositional phrases into three categories: location markers, temporal markers, and special uses (such as *on the other hand*, the second most common prepositional phrase in the data); and noun phrases are divided into the same categories with the addition of what she labels 'text markers' such as *the rest of the*. In this same

[11] Unsurprisingly, this was also one of the most common 4-word phrases found in the UGALECT Corpus.

vein, by comparing the types of N + P clusters in use by L1 and L2 writers in a general semantic taxonomy, we can focus on those that may be more problematic for learners (see Section 4.3).

Levy (2003) did a comparative study on the use of lexical bundles in professional academic writing; proficient, native speaker essay writing; and non-proficient, L1 and L2 essay writing. First, she emphasizes the notion, set forth repeatedly by Biber, Conrad, and Cortes, that lexical bundles vary by register both structurally and functionally. In conversation, most lexical bundles consist of present tense verbs, personal pronouns, and contractions, whereas in formal, academic writing, lexical bundles are usually composed of complex noun phrases, adjectives, and prepositions (Biber & Conrad, 1999). In addition, Levy (2003) observes that "bundles in conversation are generally clausal, often a pronoun followed by a verb phrase, while bundles in academic prose are phrasal, often used for physical descriptions or abstractions to mark logical or temporal relationships" (p. 33). Most often, lexical bundles are used to structure academic discourse in informational writing, while they are used to mark concrete concepts such as location and time in conversation (p. 34). Levy demonstrates that both ESL and non-proficient NS writers "have not developed the knowledge of academic vocabulary and the grammatical structures in which it occurs" (p. 1), and they frequently overuse less formal, conversational bundles inappropriately in their academic writing. Especially because of this register appropriacy issue, Levy (2003) contends that "memorized and conventionalized formulaic language is much more important than linguists believed in the past" (p. 4). Writing teachers have different expectations for word choice, both from

native and non-native speakers, and both groups tend to display an overuse of conversation conventions, especially early in their college-level curriculum.

Formal writing values "economy of expression" (Tribble & Jones, 1997, p. 59), which is very frequently achieved through the post-modification of noun phrases by prepositional and participial phrases instead of by relative clauses, where the relative pronoun and copula verb need not appear. For example, *the topic discussed at the meeting* would be considered a more sophisticated, concise writing style than *the topic that was discussed at the meeting*. And, *the book on the table* is more elegant than *the book that is on the table*, which is something we might hear from a native Spanish speaker because the use of relative clauses to post-modify nouns is more common in their L1 than in English (Moreira-Rodríguez, 2006).

Every L2 teacher has had some discussion in the classroom where meaning is not the appropriate guiding principle behind the use of a particular form. For example, when students are instructed to say *the topic in the paper*, but *the ink on the paper*, they may object that certainly the ink is *in* the paper more than the topic is *in* the paper. A frequent teacher response to this type of semantic reasoning on the part of their students is, "Well, that's just the way we say it." Corpus data offers us an accurate and objective way to empirically discover what the habits of usage are without having to rely on often fallible, intuitive guesses based on traditional, and possibly out-dated, static grammars. Learners could be satisfied with doing just what native speakers do. Language is constantly in a state of flux, and researchers exposed to a large amount of data through a corpus will be surprised by some regular patterns of usage of which they were not previously aware.

Access to large corpora now makes it possible to enlighten ourselves about the patterns of language in use rather than relying on personal intuitions.

Biber et al. (1999) demonstrates that on average there are 300 nouns per every 1000 words in academic prose and textbooks, which is more than any other group of content words. Indeed, in the UGALECT Corpus described in Section 3.5, common nouns outnumber prepositions by almost exactly two to one. Which of these nouns are commonly post-modified by prepositions will be investigated in Chapter 3.

In the next chapter, we will see some evidence of L2 writing errors with prepositions with a particular focus on those following nouns, and we will extract the N + P clusters in common usage by native speakers in writing their first-year, college compositions. Further proportional analyses (see Section 4.2) of the degree of attraction between a noun and its prepositional post-modifier will serve as robust evidence of their status as phrasal. Finally, the learner data will be checked for usage of the most frequent and robust N + P clusters from the NS Corpus.

CHAPTER 3

METHOD

The 500,000-word original corpus under analysis here was built from first-draft essays in the electronic portfolios of approximately four-hundred undergraduate students taking their first, college-level composition course at The University of Georgia in the Spring semester of 2008. Using a free, downloadable concordancer software program, AntConc 3.2.2w, created by Dr. Laurence Anthony at the University of Waseda in Japan and available at http://www.antlab.sci.waseda.ac.jp/, a word frequency list (see Appendix A) was then generated from which a list of the rank order of preposition frequencies in the essays could be determined. The immediate left collocates of the ten most frequent prepositions in the Corpus were isolated using the cluster function in the concordancer (see Appendixes B-K), and those found to be nominal colligates were then searched in order to derive a percentage of their occurrences as adjacent to particular prepositions as opposed to some other grammatical structure. Those lexical nouns having a high percentage of their occurrences with a particular prepositional right colligate not part of a separable, phrasal verb, such as *aspect(s) of, reason(s) for,* and *solution(s) to*, are then judged to be worthy of greater attention in second language writing because of their ubiquitous nature in L1 usage as demonstrated by frequency counts, proportion tests, and dispersion plots, which can visually display whether a particular form is used throughout

a corpus (hence by different language users) or is merely some common, but idiosyncratic usage prevalent in just one or few sections of the corpus.

3.1 Nature of the Study

This is a quantitative/qualitative study based on the previously referenced findings with regard to lexical bundles in academic writing. From previous studies of ESL error analysis (see Section 2.2), English language learners from many differing L1 backgrounds have demonstrated particular problems with preposition usage in their academic writing. Furthermore, the written academic register has been shown to be relatively dense with regard to the use of nouns and their cohorts, which include prepositions (see Sections 2.3 - 2.4).

This chapter will begin with the field research, which was conducted in May of 2008 for the purpose of collecting student essays from L1 Spanish speakers who were also advanced English language learners attending two different educational institutions in South America. The students' errors with regard to prepositions are first categorized qualitatively as being dependent on their immediate lexical contexts. Also, learner data with regard to prepositions following nouns is included below from the researcher's own, on-going ESOL introductory composition classes (ENGL 1101) at The University of Georgia, which have consisted of speakers of various Asian languages who are also at an advanced English language level.

We will then discuss the building of the NS Corpus, which shall be called UGALECT, and the use of a concordancer software program (AntConc) to extract examples of the most frequent N + P clusters by looking for the immediate left, nominal

colligates of the top-ten prepositions occurring in the NS Corpus. The UGALECT Corpus will also be searched for occurrences of the learner errors with regard to nouns that are post-modified by prepositions in order to objectively determine whether native speakers ever produced such specific errors.

The top-ten prepositions in the 500,000 word UGALECT Corpus (see Appendix A) with nominal left colligates occurring five times or more were recorded (see Appendixes B-K). The learner data was then searched for high-frequency, two-word N + P clusters using the concordancer in order to determine if the L2 writers were using such structures as the native speakers had. After automated part-of-speech tagging of the data, the nominal density of the writing samples was also calculated both for the learners and the native speakers by dividing the number of common nouns by the total number of words in each data set.

3.2 Primary Evidence of Learner Difficulty with English Prepositions

The field research for this project involved the collection of academic essays from native Spanish speakers in order to document their L2 errors with English preposition usage. The study was deemed exempt from UGA Internal Review Board for Human Subjects Research approval because all participants remained anonymous, their participation was voluntary, and there was no risk involved with participation in the study. No demographic information was collected on the students because the only criterion for participation in the study was that they be native Spanish speakers at an advanced L2 English level and that they had had some prior experience with academic essay writing in English. In exchange for participation, students received individual,

written feedback (provided electronically by the researcher through e-mail) on their grammar usage, essay organization, and topic development in the submitted essays.

Academic writing samples were gathered from 16 entry-level college students in an EFL teacher training program at the Universidad Andres Bello in Viña del Mar, Chile and from 32 high school seniors at the Colegio del Sol in Asunción, Paraguay.[12] Only those students 18 years of age or older participated in the study. Both groups of students had been in a secondary education program conducted entirely in English, so they were advanced level speakers with some experience in academic writing in English.

Both data-gathering sessions were carried out in exactly the same manner in a computer lab/classroom provided by the respective schools. The South American students were first presented with a workshop conducted by the researcher on the academic writing process. For approximately thirty minutes, we discussed the process of first choosing, brainstorming, and outlining a topic, and then the drafting, editing, and revision processes in order to heighten the students' awareness of writing clearly for a reader and the practice of writing multiple drafts. In their essays, the students were asked either to describe an influential person in their lives or to explain the process involved in a particular skill or hobby (recipes were disallowed). Alternatively, they could choose a topic of social significance in their respective countries from a list of general topics including, but not limited to, arranged marriage, poverty, government corruption, child labor, traditional medicine, public transportation, etc. After spending approximately twenty minutes brainstorming and outlining their individually chosen topics, the students

---

[12] The researcher wishes to thank Stael Ruffinelli de Ortiz and Juan Antonio Avalos Pinto for access to their students for this study.

then typed their essays in the computer lab for an approximate duration of one-and-one-half hours and submitted them to the researcher electronically as Microsoft Word documents in e-mail attachments. The students were allowed to use both English-English and/or Spanish-English dictionaries while typing their essays, and they had full access to the internet if they wanted to spend some time researching their topics.

Using the Track Changes feature in the word processor, the researcher then read and edited these first-draft essays remotely and sent them back to the students individually by e-mail with editing and revision comments and suggestions, which were not part of this study. The students then wrote second drafts and turned them in to their respective writing teachers for further evaluation. The original, unedited first drafts were combined and treated as one data set by the researcher, who then compiled a list of NNS errors with regard to English preposition usage below (Tables 3.1-3.4). The learner data consisted of exactly 21,483 words of running text in a total of 48 essays of approximately 400 to 500 words each.

Both native speakers and non-native speakers of a language have a range of choices with regard to prepositions in English, and non-native speakers even at advanced levels frequently choose inappropriate or unnatural-sounding ones in their spoken and written productions. In this analysis of NNS usage of English prepositions in academic writing, the following errors, as judged by the researcher, were found with regard to preposition usage. Each error is listed below along with its appropriate American English equivalent. The preposition errors were divided into four categories depending on their immediately adjacent lexical environments and on whether the preposition error could be determined by the following noun phrase alone, i.e. the object of the preposition, or it

entailed some interplay with the preceding grammatical structure, a verb, adjective, or noun, e.g. **consist in, *surrounded on,* and **interest about*. In other words, the lists are divided by the immediate structural environments of the preposition errors and whether the preceding or following environment or both of these determine the use of a particular preposition:

Table 3.1 L2 Errors in Prepositional Phrases

L1 Spanish-speaker Errors	Edited American English
in each time	*each time*
on a recent report	*in a recent report*
along the history	*throughout history*
for economic problems	*because of economic problems*
at/by the contrary	*on the contrary*
at mother's day	*on Mother's Day*
in her Confirmation	*at her Confirmation*
in the television	*on the television*
in the radio	*on the radio*
in parties	*at parties*
at their classes	*in their classes*
because of our own benefit	*for our own benefit*
in consequence	*as a consequence*
in front of a problem	*confronted* (adj.) *with a problem*
in the hill	*on the hill*
in the ticket	*on the ticket*
with a dress and heels	*in a dress and heels*
against to me	*against me*
in the coast	*on the coast*
in San Martin Avenue	*on San Martin Avenue*
near to Muelle Vergara	*near Muelle Vergara*

As can be seen in Table 3.1, the Spanish speakers exhibit confusion especially in choosing between *in* and *on* in English, which could be predicted from a contrastive analysis of what constitutes a semantic split for these students, that of the single Spanish preposition *en*. These examples also demonstrate some epenthesis of English prepositions such as in **against to me* and **in each time*.

Table 3.2 provides all of the preposition errors occurring in the Spanish-speaker essays after verbs:

Table 3.2 L2 Prepositions following Verbs

L1 Spanish-speaker Errors	Edited American English
contribute with her growth	*contribute to her growth*
discuss about	*discuss*
counted with a hand	*counted on one hand*
address to me	*address me*
fight for clothes	*fight over clothes*
ask to you	*ask you*
affects to the society	*affects society*
go on the streets	*go down the streets*
arrive to the place	*arrive at the place*
look you	*look at you*
deal up with	*deal with*
call to each one	*call each one*
stop with it	*stop it*
give to my partner	*give my partner*
count with your soulmate	*count on your soul mate*
attend to class	*attend class*
help on how to write	*help with how to write*
consist in	*consists of*
think on the topic	*think of the topic*
may sound as a fun activity	*may sound like a fun activity*
escape to my problems	*escape from my problems*

All of the preposition errors in Table 3.2 except **look you* demonstrate substitution or epenthesis errors with regard to English prepositions following verbs. For example, in the case of **fight for clothes* the student used *for* when s/he meant *over*, and another student added *to* in **affects to the society*.

Table 3.3 shows all the Spanish-speaker errors with prepositions following adjectives in English:

Table 3.3 L2 Prepositions following Adjectives

L1 Spanish-speaker Errors	Edited American English
combined to the noise	*combined with the noise*
hard to me	*hard for me*
surrounded of many people	*surrounded by many people*
stolen to	*stolen from*
driving on their cars	*driving in their cars*
passive upon something	*passive about something*
disappointed of this place	*disappointed by this place*
directed to young people	*directed at young people*
usual in first timers	*usual for first timers*
focusing in catching	*focusing on catching*
related with	*related to*
fulfilled with	*fulfilled by*
thinking in what to make	*thinking of what to make*

All of the preposition errors in Table 3.3 represent problems with substitution, i.e. using the inappropriate preposition with the preceding adjective. Also, most of these adjectives represent participial forms derived from verbs as in *combined* and *thinking*.

Table 3.4 shows all preposition errors after nouns made by the Spanish speakers:

Table 3.4 L2 Prepositions following Nouns (Adjectival Modifiers)

L1 Spanish-speaker Errors	Edited American English
problem of everyone	*problem for everyone*
poverty to the country	*poverty in the country*
corruption in children's rights	*corruption with regard to children's rights*
details of him	*details about him*
help for something	*help with something*
example for this quality	*example of this quality*
reasons of it	*reasons for it*
city from Argentina	*city in Argentina*
revenge with someone	*revenge on someone*
thing of having a sister	*thing about having a sister*
opinion in the situation	*opinion of the situation*
a look on the themes	*a look at the themes*
meaning on the usage	*meaning of the usage*
life on danger	*life in danger*
importance in control themselves	*importance of controlling themselves*
time of going to some bars	*time for going to some bars*
looking their surroundings	*looking at their surroundings*
responsibility from the one	*responsibility on the one*
interest about something	*interest in something*
decrease on the number	*decrease in the number*
programs in their computers	*programs on their computers*
effects to society	*effects on society*

For the preposition errors following nouns, the choices the students made were deemed inappropriate by the researcher due to the interplay of the preceding noun with the object of the preposition (another noun or a pronoun) and not due to the object of the preposition in and of itself. For example, there is nothing wrong with *for this quality* or *from Argentina* when considered alone. However, **example for this quality* and **city from Argentina* represent preposition errors following nouns. All of the errors in Table 3.4 except for *looking their surroundings*, which is omission, involve substitution of an inappropriate preposition for the context.

As demonstrated by the four tables above, preposition errors of all types involving substitution, omission, or epenthesis occurred in all environments. In order to check the objectivity of considering these uses inappropriate, the UGALECT Corpus was subsequently searched for any occurrence of the learner-produced phrases above. The L1 Spanish speakers provided examples of preposition usage (or non-usage) that can be regarded as distinctively non-native because all examples of the NNS errors, as judged by the researcher, were subsequently searched for in the UGALECT Corpus in order to objectively verify that NSs did not produce such contiguous sequences in a span of 500,000 words. In searching for each preposition as used by these native Spanish speakers along with its immediate left and/or right collocates (2-4 word contiguous sequences), the concordancer software returned no hits in the UGALECT Corpus, verifying that these particular phrases were not used even once by native speakers in a 500,000 word span. For example, although the contiguous sequence *of everyone* did occur three times in the NS Corpus as in *in front of everyone, the attention of everyone,* and *the safety of everyone*, and the contiguous sequence *problem of* occurred twenty times, there were no occurrences of the phrase *problem of everyone* nor *everyone's*

problem (which is what the NNS student meant) in the NS Corpus. In fact, no occurrence of *problem of* was followed by a pronoun of any sort; it was followed by a noun phrase in every instance in the UGALECT Corpus.

Because many of these preposition choices depend on a preceding noun phrase and because academic/informational writing has been demonstrated to be nominally dense (see Section 2.4), the decision was made to focus on NS usage of prepositional phrases functioning adjectivally as post-modifiers of nouns. This decision was also made in light of the fact that there is already coverage of prepositional phrases and multi-word verbs and adjectives in current ESL textbooks (see Section 1.1). However, to the researcher's knowledge, there is no coverage of N + P clusters as viable lexical units in extant ESL teaching materials.

As further evidence of learner errors with English preposition usage, examples of erroneous usage or non-usage of prepositions after nouns in L2 English academic writing were also recorded from the academic essays of native speakers of various Asian languages including Korean, Chinese, Japanese, and Vietnamese students, who were taking ESOL freshman composition courses with the researcher as instructor at The University of Georgia in the 2007-2008 terms. Most non-native speakers admitted to the University attended high school in the U.S. and are frequently referred to in the literature as 'Generation 1.5,' meaning they immigrated to this country with their parents, who were not born in the U.S. They are bilingual with some residual, possibly fossilized, usage errors evident in their academic writing, including errors with English prepositions. The examples of preposition errors following nouns in Table 3.5 were extracted from the first-draft essays submitted by these students in their electronic portfolios for ENGL 1101:

Table 3.5 Asian Learners' Preposition Errors following Nouns

Preposition Errors	Edited American English
admission in UGA	*admission to UGA*
scholarships about music	*scholarships in/for music*
reasons on that	*reasons for that*
the mean being	*the meaning of being*
one day hard work	*one day of hard work*
earphones on their ears	*earphones in their ears*
hints on their music	*hints in their music*
a big role of music	*a big role in music*
details on a travel	*details about the trip*
decision for the place	*decision on/about the place*
the thought it	*the thought of it*
basic skills on math	*basic skills in math*
a period time	*a period of time*
adjustment kindergarten	*adjustment to kindergarten*
a key helping	*a key to helping*
a reaction the situation	*a reaction to the situation*
hundreds years ago	*hundreds of years ago*
the demand the students	*the demand on the students*
lifestyle the politicians	*lifestyle of the politicians*
an article of newspapers	*an article in the newspapers*
inconvenience for these things	*inconvenience of these things*
a few pages newspaper	*a few pages of the newspaper*
the penalty of cheating	*the penalty for cheating*
help for homework	*help with homework*
revenge the allies	*revenge on the allies*
performances on sports	*performance in sports*
influences to students	*influences on students*
thousands miles away	*thousands of miles away*
attention on the children	*attention to the children*
understanding to freedom	*understanding of freedom*

The errors in Table 3.5 demonstrate that English prepositions following nouns are also a challenge for speakers of various Asian languages. The examples from these learners represent a greater rate of error with regard to English prepositions (14%) than do the Spanish-speaker errors (10%). A qualitative consideration of the types of mistakes reveals a greater incidence of omission in the Asian students' productions, and the Asian students rarely epenthesized English prepositions as the Spanish-speaking participants had, most often with regard to *to*, which usually translates as Spanish *a*.

3.3 Demographics of the NS Participants

The University of Georgia admits approximately five-thousand incoming freshmen per academic year, all of whom must take or exempt the two, first-year writing courses, a common requisite at many U.S. colleges and universities (Desmet, personal communication). Because this study was conducted anonymously, no identifying characteristics of the individual writers were saved. A general demographic of incoming freshmen for the 2007-2008 academic year can be obtained from the undergraduate admissions office website at http://www.admissions.uga.edu/4_fy_closerlook.html.

Non-native English speakers attending the University are held to the same rigorous standards as native speakers; they are required to take the Scholastic Aptitude Test (SAT) and to submit high school Grade Point Averages (GPAs). However, non-native speakers have the option of taking first-year composition classes specially designed for ESOL students. Those classes require permission (POD) of the First-Year Composition (FYC) office in order to register, and none of the essays from those designated ESOL sections (as could be determined by the individual instructor listed for each course) were accessed for this study, which aims for a descriptive analysis of native-speaker usage.

As noted at the above referenced website, the 2007 entering UGA freshman class consisted of 63% females, and 20% of the freshman class was non-Caucasian. Eighty-three percent were Georgia residents from 400 different high schools and 144 different counties in Georgia. The average SAT score for entering freshmen in 2007 was 1233 with an average high school GPA of 3.79 (http://www.uga.edu/profile/facts.html). Therefore, the UGALECT Corpus is meant to be representative of the academic writing of this student populace.

3.4 The <emma> Archive

Freshman composition teaching has evolved quite a bit over the past few decades. Today, the process approach to writing allows students the opportunity to polish their writing with teacher input and a greater focus on learning how to improve their crafting of exposition and argumentation. Students submit their documents electronically in a serial exchange with their instructors and peer reviewers. As readers of peer work themselves, students also develop a greater appreciation of writing clearly for a reader.

<emma> is an electronic mark-up and management application that allows for the archiving of written drafts from students in composition courses. One of its primary purposes is to allow for interactivity in electronically-stored text documents both between the composition instructor and the student writers and among students for peer review. Another advantage of archiving student compositions is that there is a permanent record of all draft submissions, in this case since 2002 at UGA, allowing the students to build a comprehensive portfolio of their writing progression throughout the semester, which also encourages the students to focus on writing as a process of editing and revision. Of course, the ultimate advantage for researchers is the archiving of an expansive amount of data available for analysis (Desmet & Balthazor, 2005). Upon creation of an <emma> account, students are asked if they will allow their submissions to be accessed for research purposes. Consequently, only work by those students having granted permission in advance is accessible to researchers.

First-year composition students at The University of Georgia are instructed to set up a web-based account on the <emma> homepage, where they can store and manage all drafts produced during the semester in separate folders. A final portfolio consisting of (a) a brief biography, (b) an introductory reflective essay, (c) two polished, final draft essays,

(d) a revision exhibit, (e) a peer review exhibit, and (f) a "wild card" exhibit is then submitted at the end of the semester for partial consideration in their final grade for the course. Upon initiating their account, each student is asked for permission to use their written work in research conducted under the auspices of the First-Year Composition Office. The essays of those students who did not give permission to store their work for future research purposes are not permanently archived and cannot be accessed by anyone but their instructor and fellow students (as permitted peer reviewers) during the course of the semester.

The Open Office word processing software used in conjunction with <emma> is designed to allow for such collaborative writing and uses the .odt format for documents produced for uploading to the <emma> archive. Incorporated in this program is a commenting function, which can be utilized by both instructors and students in evaluating rhetorical style and grammatical usage. Of course, the extent of utilization of this particular feature is up to the discretion of each instructor, and some instructors elect to use a word processing program they are already more familiar with such as Microsoft Word. Both .odt and .doc formatted essays were copied-and-pasted to the UGALECT Corpus for use in this study. The complete file was then saved as one Word document, which was subsequently converted to a plain text document in Notepad (2.80 MB) as required by the concordancer because complex formatting can interfere with the operation of the software.

3.5 Building the UGALECT Corpus: Data Transformation

The UGALECT Corpus is meant to represent the writing habits of native speakers at the beginning of their college careers and was analyzed for the purposes of this study

with regard to NS usage of N + P clusters (N + P). Such N + P clusters were isolated by first targeting the most frequent prepositions occurring in the Corpus, which consists of approximately 600 first-draft essays from 15 different sections of ENGL 1101 from the Spring semester of 2008.[13] A cutoff was made at exactly 500,000 words of text after being edited for spelling, typing and punctuation anomalies that could have affected word count frequencies. For example, some students were in the habit of leaving a space on either side of periods, which would result in the word processor counting a period as a word, which are after all just a series of characters between two white spaces for the software program. Therefore, those spaces were manually deleted throughout the entire Corpus by using the FIND and REPLACE (Control-F) commands in Word in order to get a more accurate word frequency count.

One of the many advantages of keeping an electronic database of student essays is that it allows for the extrapolation of specific document features such as thesis statements or of labeled folders of various submissions such as first-draft essays. Only essays in first-draft folders were accessed for this study although the researcher makes no claim for accuracy in this regard because sometimes students did misfile their submissions. For example, some outlines and journal and/or biographical entries were found in a few of the first-draft folders; however, such submissions, which were obviously not first-draft essays, were not copied to the UGALECT Corpus.

The sampling of essays for this study was not completely random for several reasons. Only essays filed as first drafts by the students were considered for copying to

---

[13] The researcher wishes to thank the head of the First-Year Composition Office at UGA, Dr. Christy Desmet, for permission to access the electronic archive of freshman essays for use in this study.

the Corpus in order to avoid any teacher input such as editing or revision advice and to minimize the amount of quoted, outside, or other language from research sources. In order to maximize the frequency of nominal phrases, narrative writing, which tends to use more lexical verbs (Reid & Byrd, 1998), was not included in the Corpus. To keep idiosyncratic usage by any individual writer from affecting the word frequency counts (Biber, Conrad, & Cortes, 2003), no more than two essays from any one student's portfolio were copied to the Corpus. To avoid retaining identifying information, prose judged to be of a personal nature, such as autobiographical narratives or journal entries, was not copied to the Corpus. Other text types not copied to the Corpus were lists, outlines, travel descriptions, past experience narratives, reflective (having to do with the writing process) essays, and any peer reviews or revisions that had been misfiled in a student's first drafts folder. To capture a more formal register such as that characteristic of academic rhetoric, only essays of an argumentative or expository nature were retained. This would include letters to the editor, political opinion essays, literary descriptions or critiques, responses to visual imagery, process descriptions, argumentative essays, and so forth.

Using the AntConc concordancer to take a look at dispersion plots of selected items also helped to guard against anomalous frequencies that may be characteristic of a particular group of writers or a particular topic assignment. A quick glance at a dispersion plot in the concordancer can display the distribution of a lexical item or phrase throughout the entire Corpus with black vertical bars along a horizontal axis representing every occurrence of a particular search term or phrase. Such 'local repetitions' can be due to "immediate topical concerns of the discourse" (Biber, Conrad, & Cortes, 2003), but prepositions are almost always evenly distributed in a corpus of running text because

of their vast utility in joining the more lexical units of discourse and the fact that they are a closed word class, somewhat limiting variation in usage.

The length of each sample essay ranged between about 700 to 1500 words. The approximate number of individual compositions was 600 from a total of 15 different English 1101 classes, although a few of the classes had the same instructor, which was evidenced by the topics covered having some effect on the repetition of particular common and especially proper, i.e. capitalized, nouns.

All citation information, such as works cited lists and in-text citations, was either not retained or later deleted. Utilizing the FIND command in Word, all parenthetical information in the entire Corpus was reviewed. If the information was in the form of running text, it was retained, and if information within parentheses consisted of a name, date, page number, and/or abbreviations, it was deleted so as not to influence the total word count. However, it was decided to retain all quotations because these were likely to be in the appropriate academic register. All website addresses were also located and deleted using the FIND command. All formatting such as boldface, italics, and underlining was removed. The text was finally saved in Tahoma font, size 10 with very narrow margins for a total of 464 pages of running text in Word.

Frequent items in a corpus such as prepositions tend to be more stable in their distribution (Biber, 1988), i.e. more evenly distributed than less frequent items. Such distribution for particular words and/or phrases in the corpus can be checked by a quick glance at a dispersion plot of the selected item as provided by the AntConc concordancer software program. A cutoff of the texts collected was made at an even 500,000 words in

the UGALECT Corpus because this is a common word count in many of the extant, mid-sized, non-monitor corpora.[14]

Tagging of a corpus can be done manually, which is extremely time-consuming, or by using an automatic tagger such as CLAWS (Constituent Likelihood Automatic Word-tagging System) available on-line for license purchase from Dr. Paul Rayson[15] at http://ucrel.lancs.ac.uk/claws/ at the University Centre for Computer Corpus Research on Language in Lancaster, England. However, automatic taggers are not error free, and the best accuracy rate is 96-97% with the relatively accurate CLAWS POS tagger. In a mid-sized (by today's standards) corpus such as UGALECT with 500,000 words, this could potentially produce up to 20,000 lexical items incorrectly tagged for part of speech. In order to verify automated tagging, the immediate co-text of all prepositions in the corpus was checked manually for accuracy.

The cutoff of twenty occurrences of N + P clusters per 500,000 words was set prior to any analysis based on the precedent of twenty occurrences per million words for 4-word lexical bundles (Biber et al., 1999). The structural unit of analysis for this study is N + P, where the nouns are lexical (i.e. not proper names) and could appear in their singular, plural, or non-count forms. The total number of common nouns in UGALECT is 114,075 (23%), and the total number of prepositions including *of* is 58,239 (11.6%), which is in-line with previous findings for formal, written English (Kennedy, 2003).

---

[14] A monitor corpus is one that is continually being added to as a diachronic record of language in use.

[15] The researcher wishes to thank Dr. Paul Rayson of the UCREL at Lancaster University, UK for assisting with the tagging of the 500,000-word UGALECT Corpus for this study.

3.6 The AntConc Concordancer and CLAWS5 POS Tagger

The entire UGALECT Corpus was initially saved as a Microsoft Word document in order to utilize the features of the word processor as described above. The text was then saved as a plain text document in Notepad for processing through the concordancer, AntConc, a free, downloadable software program for use with corpora, available from Dr. Laurence Anthony's homepage at Waseda University in Japan (http://www.antlab.sci.waseda.ac.jp/software.html). Non-formatted text is a general requirement for use with any concordancer so as not to interfere with the operation of the software. An extensive corpus analysis was conducted, beginning with the list of absolute word frequencies for the entire 500,000 word corpus (see Appendix A for the first 600 words).

As determined from the concordancer-generated list of word frequencies, the thirty, most common words possibly functioning as prepositions along with their rank and raw word frequency from Appendix A are listed below in Table 3.6:

Table 3.6 Absolute Word Rank and Frequencies for Potential Prepositions

RANK	FREQUENCY	TOKEN	RANK	FREQUENCY	TOKEN
2nd	16295	*to* (inf/prep)	96th	598	*over* (prep/adv)
3rd	14742	*of*	98th	594	*through* (prep/adv)
6th	9852	*in* (prep/adv)	117th	491	*after* (prep/sub)
9th	4635	*for* (prep/conj)	144th	398	*between*
12th	3680	*as* (sub/adv/prep)	161st	364	*around* (prep/adv)
17th	3463	*with* (prep/adv)	169th	346	*before* (sub/prep)
18th	3456	*on* (prep/adv)	171st	346	*without* (prep/adv)
27th	2043	*by* (prep/adv)	178th	325	*since* (sub/prep)
31st	1880	*from* (prep/adv)	180th	321	*during*
42nd	1456	*at* (prep/adv)	209th	266	*against*
51st	1309	*about* (prep/adv)	228th	243	*off* (adv/prep)
65th	977	*like* (v/prep/adv/adj)	231st	240	*down* (adv/prep)
67th	957	*out* (prep/adv)	307th	189	*throughout*
75th	868	*up* (adv/prep)	309th	188	*toward(s)*
85th	743	*into*	321st	183	*within* (prep/adv)

As noted in the chart, some words can function as either prepositions or other word classes, and this function can be determined by checking their immediate or extended context in the concordance lines. For example, the left context can usually distinguish between a particle and a preposition,[16] with particles functioning as parts of phrasal verbs, as in *put out*, and prepositions functioning as complements to verbs, nouns or adjectives, as in *abide by, interest in* and *afraid of*. A prepositional phrase functioning as a complement to a verb phrase or an adjective would be considered an adverbial and, therefore, not relevant to this study of adjectival prepositions. The present study seeks to isolate only those prepositional phrases functioning adjectivally, i.e. as complements to nouns. So, the longer left span in the line of text (i.e. the concordance) also had to be checked to ensure that any given preposition was not part of a separable phrasal or prepositional verb such as in *let the truth out*. In other words, the prepositions in such cases would not be functioning adjectivally and, therefore, were eliminated from consideration in the calculations for N + P clusters.[17]

The UGALECT Corpus was tagged using the CLAWS5 POS on-line tagger, which utilizes a 62-category tag set for parts of speech.[18] Total word counts were taken from the initial word frequency list (Appendix A) generated by the concordancer, and the

---

[16] For a detailed analysis of the finer distinctions among particles and prepositions following verbs, see O'Dowd (1998). For historical distinctions in the development of phrasal and prepositional verbs, see Brinton and Traugott (2005). For a quantitative analysis of phrasal and prepositional verbs, see Biber et al. (1999).

[17] It should be noted here that quite often, as in the case of *of*, prepositional phrases also function adjectivally as modifiers of pronouns (see Appendixes C & L), another potential ESL/EFL teaching point, though not included in this study.

[18] The CLAWS POS Tagger has been expanded to tag much finer distinctions among parts of speech, but the CLAWS5 POS Tagger used here was considered adequate for this analysis of prepositions. The CLAWS5 Tagger does distinguish *to* when used as a preposition from *to* used as an infinitive marker.

number of occurrences as prepositions was taken from the tagged text. Where there is one number in Table 3.7 below, the item was always tagged as a preposition by the CLAWS5 POS tagger. Where there are two numbers, the first is the number of occurrences of that item tagged as a preposition, and the second is the number of raw occurrences of that particular item in the Corpus.

Table 3.7 Number of Preposition Occurrences/Total Word Occurrences

of	14,742	***during***	321	***inside***	55
in	9399/9852	***after***	320/491	***regarding***	54
to	6251/16,295	***against***	266	***above***	51
for	4395/4635	***around***	200/364	***per***	49
with	3459/3463	***toward(s)***	188	***since***	41/325
on	3282/3456	***throughout***	187/189	***concerning***	38
by	2005/2043	***within***	179/183	***near***	38
from	1871/1880	***before***	177/346	***down***	34/240
as	1643/3680	***under***	129/145	***onto***	32
at	1307/1456	***along***	128/143	***beyond***	29
about	1134/1309	***behind***	105/126	***until***	20/155
out	779/957	***upon***	104	***below***	14/20
into	743	***among***	102	***except***	12/25
like	735/977	***off***	82/243	***underneath***	5/6
through	544/594	***across***	78/110	***beneath***	4
between	390/398	***up***	78/868	***amid***	2
over	368/598	***despite***	74	***beside***	2
without	341/346	***outside***	58/102	***till***	1/4

As can be seen in this chart, *of* always functions as a preposition, whereas *to* usually functions as an infinitive marker, not a preposition, moving it to third most frequent preposition rather than first. With regard to most of the other top-ten prepositions (*in, for, with, on, by, from, at, about*), they almost always function as prepositions rather than as adverbials as in *hand in* and *take on*. As for tagging errors, in 11 occurrences *for* was tagged as a subordinator by the tagger, when it was actually functioning as a coordinating conjunction; its semantically-equivalent subordinator, *because*, was much more common between clausal elements. The term *as* was tagged as

an adverb in 446 occurrences, and as a preposition in 1643; however, many of its prepositional functions were in multi-word prepositional constructions, i.e. *such as* (338x), *as well as* (97x), *as opposed to* (16x), *as for* (8x), and so forth, which were not considered further. In other words, *as* operated most often as a prepositional component, very often as an adverbial subordinator, and very rarely as a noun complement, so it was disregarded from the list of the top-ten, one-word prepositions, and *about* replaced it for consideration of its occurrences as a right colligate of nouns (see Appendix K).

Prepositions are almost always followed by a noun phrase, except in the case of clause-final or so-called 'stranded' prepositions, e.g. *What's it made of?*, which occur much more frequently in conversation, usually at the end of *wh*- questions, than in academic prose. In fact, such clause-final prepositions are said to be characteristic of more involved, interactional forms of discourse such as conversation (Biber, Conrad, & Reppen, 1998, p. 148). Thus, in a corpus of academic writing, we can expect to find more noun phrases, i.e. nouns and their attendant determiners and/or attributive adjective(s), as the immediate right colligates of many prepositions, i.e. prepositional phrases.

The present study determines the nominal left colligates of the ten most common prepositions in the corpus and builds from there by looking for pattern frequencies with the resulting most common two-word sequences (N + P) recording significant findings along the way. Biber et al. (1999) used a cutoff of 20 tokens per million for determining frequent four-word lexical bundles, so this study applies an initial cutoff of 20 two-word tokens per 500,000 words for further consideration as N + P clusters.

3.7 Procedural Considerations: Prepositions as Other Word Classes

As demonstrated above, some very common words regarded as essentially prepositions can be relegated to other word classes such as particles, adverbs, coordinators, and the *to* (+ V) infinitive marker, depending on their respective contexts. Some automatic part-of-speech taggers, such as CLAWS 5 and 7, treat *of* and the *to* infinitive marker as distinctive categories with separate tags. In fact, Sinclair (1991b) feels that *of* should be treated as a distinct word class in and of itself because of its relatively large range of application and its various nuances of meaning suggesting the label 'partitive particle' for *of* instead. He contends that the main role of *of* is to combine "with preceding nouns to produce elaborations of the nominal group" (p. 83). So, again *of* is more 'sensitive to' what precedes it rather than to what follows (Kennedy, 2003; Lewis, 2000; Sinclair, 1991b).

When immediately followed by a verb or an adverb (in the case of prescriptively prohibited 'split infinitives'), *to* functions as an infinitive marker. A majority (62%) of the occurrences of *to* were found to be infinitival in the UGALECT Corpus, removing it as the top contender for preposition frequency (see Table 3.7 above).

In turn, *of* was found to be the most frequent preposition in the UGALECT Corpus, which was to be expected based on results from other English corpus studies (Francis, Kučera, & Mackie, 1982; Fries & Traver, 1950; Leech, Rayson, & Wilson, 2001). *Of* is consistently the most frequently occurring preposition in English, especially in written discourse where its many, more abstract meanings and its most common use as post-modifier of a noun can be fully exploited.

As a preposition, *for* will be followed by a noun phrase; as a coordinating conjunction, it should be both preceded and followed by clausal elements, i.e. a noun

phrase subject plus an associated inflected verb phrase as in …*discipline is acceptable for the child, for it lets the child understand*…. This goes for other prepositions that may also function as subordinating conjunctions such as *after*, *before*, *since*, and *until* as well. So, discerning prepositional usage for these particular words requires a greater span of text, which can be done by checking the individual concordance lines[19] with a span of at least 5 words to the left and right of the item in question.

Words in the top-thirty list that can be used as either prepositions or adverbs include *in, on, as*, *like, out, up, into, over, through*, *off,* and *down*. As adverbs, all of these terms may occur frequently as complements to verbs, e.g. *look out* and *give up*.

3.8 Prepositional *To*

The immediate right contexts of *to* had to be reviewed manually through the concordancer for determination of its status as a prepositional colligate to a noun in each case. First of all, the immediate right collocates of *to* were isolated using the cluster function in the concordancer (see Appendix B for those clusters occurring at least 10 times or more). The cluster function generates an ordered list of contiguous sequences that appear around a search term or phrase in the target files, in this case the UGALECT Corpus. For example, *to the* was the most common cluster having *to* as the left collocate followed by *to be*, *to a*, *to make*, and *to do*. These very frequent two-word phrases also demonstrate the more common use of *to* as an infinitive marker, i.e. as left colligate to a verb.

---

[19] The concordancer in use here, AntConc, currently does not accept annotated text, so it could not be used to search for particular part-of-speech tags in conjunction with particular words. The tagged text was searched using the Control-F command in Word, which also provides counts for searched terms with tags.

In any case where *to* was not followed immediately by a verb or an adverb (in other words, when it functioned as a true preposition), the immediate left collocates were then determined manually by looking at the individual concordance lines, and those functioning as nouns were recorded. For example, *to the*, occurring 1462 times in the Corpus, was searched as a phrase and then sorted alphabetically in order to discern nominal left colligates more easily, while *to be* (occurring 1203 times) was discarded from further analysis because it is an infinitive. The twenty most common left collocates of *to the* and their frequencies were found to be:

according (to the)	75x	*related (to the)*	16x
due (to the)	75x	*solution (to the)*	15x
appeal (to the)	32x	*go (to the)*	14x
up (to the)	25x	*relate (to the)*	13x
come (to the)	19x	*similar (to the)*	13x
back (to the)	18x	*appealing (to the)*	11x
compared (to the)	18x	*access (to the)*	9x
led (to the)	18x	*appeals (to the)*	9x
attention (to the)	16x	*close (to the)*	9x
it (to the)	16x	*comes (to the)*	9x

As can be gleaned from this brief list, there are seven words potentially being used as nouns preceding the two-word cluster in the top twenty occurrences of *to the*: *appeal, back, attention, solution, appealing, access,* and *appeals*. All contexts were then checked using the concordance list function in order to determine nominal status for these words and all other potential nouns in the longer list of those *to* collocates occurring five or more times in the UGALECT Corpus as presented in Appendix B. Only 3 occurrences of *appeal to the*, 3 occurrences of *appealing to the*, and no occurrences of *appeals to the* were found to be functioning nominally. No occurrences of *back to the* were found to be nominal as in, for example, *turned her back to the audience*, and all occurrences of *attention to the, solution to the,* and *access to the* were, of course, nouns post-modified by

a prepositional phrase beginning with *to*, which were recorded as such. This process was repeated over and over again so that all occurrences could be recorded in a list of the most frequently occurring nouns followed by prepositional *to*.

All occurrences of *to* followed by a noun, a determiner, or a pronoun (in other words, functioning as a true preposition) were searched in this same manner using the cluster function, and the frequencies of nouns, both singular and plural forms, followed by prepositional *to* were recorded. Those nominal left colligates of prepositional *to* occurring more than once in the Corpus are also listed in Appendix B.

By searching for the frequencies of each of these noun plus prepositional *to* clusters, a total number of occurrences could be determined for both singular and plural forms of the noun. All prepositions immediately adjacent to nouns also had to be checked for whether they were actually particles in a separable, phrasal verb with the noun serving as direct object to the verb, in which cases, these were discarded as not candidates for N + P cluster status.

The next step was to check each occurrence of the nouns followed by *to* only and to individually verify each as a noun followed by a prepositional *to* for a total count of this structure. The most common N + P clusters, those occurring twenty times or more as followed by prepositional *to* in the Corpus, were: *access to* (39 tokens), *solution(s) to* (39 tokens), *attention to* (30 tokens), *response(s) to* (27 tokens), *addition to* (26 tokens), *answer(s) to* (23 tokens), and *way(s) to* (23 tokens).

3.9 Nominal Left Colligates of *Of*

As has been repeatedly determined by corpus studies, *of* is the most common preposition in the English language (see Section 2.3). It always functions as a preposition

and frequently serves to connect one noun phrase to another as an adjectival complement. The top one-hundred nominal left colligates (not including pronouns) with *of* as an adjectival complement, i.e. those occurring twenty times or more in the UGALECT Corpus, are listed below from Appendix C:

part of	*age of*	*top of*	*attention of*
use of	*millions of*	*world of*	*definition of*
amount of	*purpose of*	*period of*	*freedom of*
type of	*time of*	*development of*	*future of*
number of	*importance of*	*style of*	*images of*
lot of	*state of*	*thought of*	*list of*
form of	*beginning of*	*understanding of*	*middle of*
idea of	*examples of*	*cause of*	*story of*
types of	*front of*	*control of*	*fear of*
kind of	*side of*	*loss of*	*generation of*
way of	*effects of*	*quality of*	*knowledge of*
people of	*hundreds of*	*risk of*	*meaning of*
lack of	*States of*	*terms of*	*picture of*
sense of	*years of*	*amounts of*	*pictures of*
majority of	*forms of*	*citizens of*	*population of*
aspects of	*issue of*	*course of*	*production of*
result of	*parts of*	*means of*	*death of*
University of	*point of*	*chance of*	*hopes of*
weapons of	*variety of*	*hours of*	*method of*
end of	*life of*	*center of*	*nature of*
aspect of	*sort of*	*creation of*	*couple of*
lives of	*source of*	*half of*	*level of*
percent of	*view of*	*process of*	*problem of*
thousands of	*history of*	*benefits of*	*role of*
group of	*image of*	*case of*	
rest of	*piece of*	*goal of*	
example of	*appearance of*	*ideas of*	

A concordance search of *of* proves to be quite fruitful indeed. The concordancer lists the most frequent form of a noun occurring with *of* immediately to its right, whether that form is singular or plural, capitalized or in lower case. The concordancer can also be set to disregard case and to list both singular and plural forms together using the wildcard settings. However, doing a search for *part of* in both its singular and plural forms together using the wildcard setting function for the plural inflectional ending will yield concordances for *party of* as well. So, in order to ensure accuracy, the different forms of each nominal colligate above were searched for separately. For example, *part of* occurs

191 times, *Part of* occurs 7 times, *parts of* occurs 36 times, and *Parts of* occurs once for a total of 235 times for this N + P cluster. In addition, several students rendered the phrase *a part of* as *apart of* for an additional 12 occurrences bringing the actual total for this most common N + P cluster to 247. Also, forms such as *thought of* had to be checked in all occurrences (30 tokens) for possible status as multi-word verbs. *Thought of* was found to be verbal (in some cases used as an adjective) 17 times and nominal only 13 times, and *thoughts of* was, of course, nominal in every instance (7 tokens) for a total of 20 occurrences of *thought(s) of* as an N + P cluster. So the phrase, *thought(s) of* is included in the list below having just passed the pre-determined cutoff of 20 times per 500,000 words for N + P clusters. Another case in point, *care of*, was found to be some form of *take care of*, a phrasal verb, in all thirty occurrences in the Corpus. The concordancer facilitates such searches by allowing the sorting of concordance lines alphabetically by adjacent left and/or right collocates. The thirty concordance lines for *care of* arranged alphabetically by first, second, and third left collocates are displayed in Figure 3.1 below:

begin **taking** better care of the environment. We
ould **take** life long care of the individual.
ng to keep and **take** care of their child. Howeve
e feels he can **take** care of him self and surviv
onment. We can **take** care of the planet better b
o their farms, **take** care of their houses, and a
olice officers **take** care of us; so why don't we
r if she can't **take** care of herself, let alone
not be able to **take** care of their children. Bet
re expected to **take** care of the "house work". A
ou are left to **take** care of your two siblings.
s per month to **take** care of, and how an unplann
hey are unable **take** care of their baby, then so
o why don't we **take** care of them? The city says
sured. He will **take** care of the environment by
ack Obama will **take** care of this problem. This
President will **take** care of these problems and
she could be **taken** care of by the troops. Sad
t needs to be **taken** care of as soon as possible
t needs to be **taken** care of in our society, is
oys are being **taken** care of. The real issue her
r system that **takes** care of their wants and nee
hey go about **taking** care of their clients. In t
ommunity and **taking** care of his family. His aud
ole includes **taking** care of the household, work
he can start **taking** care of the lives of their
, but values **taking** care of him or herself. Ho
tayed home and **took** care of the domestic duties
tion, his aunt **took** care of him until his mothe
The old woman **took** care of the linen. Everyone

Figure 3.1 The 30 Concordance Lines for *care of*

Because *care of* appears in the UGALECT Corpus as always preceded by some form of *take*, it was considered part of the contiguous collocation *take care of*, i.e. a phrasal verb, and not as an N + P cluster per se. So, each N + P cluster from the list above was checked for context in the concordance lines in order to determine its consistent phrasal boundaries. Any N + P cluster found to be part of a greater lexical context with relative consistency was removed from further consideration as an N + P cluster. For instance, *touch with*, which occurred twenty-six times, was found to collocate with *in…* in all of its occurrences in the Corpus and with *keep in…* and *stay in…* in 62% and 27% of those respectively. Therefore, it would be better treated as a phrasal verb.

By checking the concordancer for other forms of the noun, singular or plural, which may or may not be included in the list of the most frequent above, the number of occurrences for the lexeme may increase. A more accurate portrayal includes both singular and plural forms of the nouns occurring with *of* and their total number of occurrences. Capitalized nouns such as those beginning sentences were also included in the counts even though these were counted separately by the cluster function in the concordancer. However, proper forms, which were also capitalized, were considered highly-topical as portions of titles or names and thus were not included in this count of the most useful N + P clusters with *of* in the Corpus:

247x *part(s) of*
224x *type(s) of*
185x *use(s) of*
180x *amount(s) of*
140x *number(s) of*
133x *aspect(s) of*
130x *form(s) of*
128x *lot(s) of*
116x *idea(s) of*
108x *kind(s) of*
107x *way(s) of*
100x *example(s) of*
99x *life/lives of*
84x *people of*
81x *lack of*
78x *sense of*
77x *result(s) of*
74x *group(s) of*
74x *majority of*
68x *weapon(s) of*
64x *end of*
60x *percent of*
60x *thousands of*
59x *age(s) of*
59x *image(s) of*
58x *rest of*
58x *time(s) of*
57x *effect(s) of*
55x *purpose(s) of*
49x *state(s) of*
49x *year(s) of*
48x *millions of*
48x *side(s) of*
46x *point(s) of*
45x *beginning(s) of*
45x *history of*
45x *picture(s) of*
45x *piece(s) of*
44x *importance of*
44x *source(s) of*
43x *case(s) of*
43x *front of*
43x *sort(s) of*
42x *issue(s) of*
42x *view(s) of*
39x *style(s) of*
38x *chance(s) of*
37x *hundreds of*
37x *variety/ies of*
36x *cause(s) of*
36x *period(s) of*
35x *appearance(s) of*
35x *method(s) of*
35x *risk(s) of*
34x *act(s) of*
34x *member(s) of*
33x *top(s) of*
32x *control(s) of*

32x *level(s) of*
32x *quality/ies of*
32x *story/ies of*
32x *world of*
31x *death(s) of*
31x *problem(s) of*
30x *citizen(s) of*
30x *development of*
30x *understanding of*
29x *feeling(s) of*
29x *loss of*
29x *process(es) of*
29x *term(s) of*
28x *president of*
28x *benefit(s) of*
27x *course of*
27x *day(s) of*
27x *favor(s) of*
27x *goal(s) of*
27x *means of*
26x *fear(s) of*
26x *hope(s) of*
26x *hours of*
26x *creation(s) of*
26x *meaning(s) of*
26x *role(s) of*
26x *word(s) of*
25x *center of*
25x *generation(s) of*
25x *half of*
25x *list(s) of*
24x *freedom(s) of*
23x *area(s) of*
23x *attention of*
23x *color(s) of*
23x *cost(s) of*
23x *debate(s) of*
23x *definition of*
23x *future of*
23x *middle of*
23x *opinion(s) of*
23x *population(s) of*
22x *danger(s) of*
22x *knowledge of*
22x *need(s) of*
22x *pound(s) of*
22x *production of*
22x *woman/en of*
21x *advantage(s) of*
21x *content(s) of*
21x *couple(s) of*
21x *name(s) of*
21x *nature of*
21x *sign(s) of*
20x *city/ies of*
20x *leader(s) of*
20x *message(s) of*
20x *principle(s) of*
20x *set(s) of*
20x *thought(s)* of

3.10 Nominal Left Colligates of *In*

Functioning as a preposition in 95% of its occurrences (9399/9852), *in* was relatively easy to isolate with nominal left colligates. For some collocates such as *result* and *work*, which could be verbs, their status as nouns had to be checked in each individual context for an accurate count of true nouns. A list of the most frequent left collocates of *in* was derived using the cluster function in the concordancer. The nominal left colligates of *in* occurring twenty times or more in the Corpus were:

121x	*change(s) in*	32x	*difference(s) in*
110x	*war(s) in*	32x	*problem(s) in*
99x	*people in*	27x	*issue(s) in*
75x	*women/woman in*	27x	*student(s) in*
61x	*role(s) in*	27x	*thing(s) in*
54x	*increase(s) in*	26x	*men/man in*
40x	*time(s) in*	22x	*school(s) in*
36x	*child(ren) in*	22x	*situation(s) in*
36x	*place(s) in*	21x	*country/ies in*
35x	*life/lives in*	20x	*character(s) in*
34x	*interest(s) in*	20x	*day(s) in*
33x	*point(s) in*	20x	*debate(s) in*
33x	*way(s) in*	20x	*technology/ies in*

For all left collocates of *in* occurring five or more times in the UGALECT Corpus, see Appendix D.

3.11 Prepositional *For* with Nominal Left Colligates

For can function as a preposition or much less often as a coordinating conjunction, so greater clause-level contexts had to be checked in the concordance lines. Also, verbal colligates such as *looking*, *fighting*, and *searching* had to be determined to be functioning as gerunds, e.g. *Searching for answers is time-consuming*, in which case they are included as nouns collocating with *for*, or as participial verbs or adjectives, e.g. *They are fighting for a cause*, in which case they are not. The nominal left colligates of *for* occurring more than twenty times in the Corpus are:

77x	*reason(s) for*	28x	*time(s) for*
34x	*need for*	23x	*plan(s) for*
29x	*order for*	23x	*room for*
28x	*life/lives for*	22x	*candidate(s) for*

For all left collocates of *for* occurring five or more times in the Corpus, see Appendix E.

3.12 Prepositional *As*

As tagged by the CLAWS5 POS tagger, very many occurrences of *as* in the UGALECT Corpus were as subordinating conjunctions beginning clausal elements, which contain a subject noun phrase and a finite verb phrase, for instance, *As we look at our own community,….* For *as* to be functioning as a true preposition, it would have to be followed by a nominal with no associated inflected verb, i.e. not a clause. The occurrence of *as* as a subordinator totaled 45%, almost half of all occurrences (1654/3680), according to the automatic tagger, and the occurrence of *as* as a preposition, for example in *as a matter of fact* or *as a result*, was approximately the same, 45% (1643/3680). Such common prepositional phrases beginning with *as* never function as complements to nouns. Indeed, most of the 1,643 occurrences of *as* functioning as a preposition were actually parts of adverbial prepositional phrases associated with a preceding adjective or verb phrase such as in *…the friendship he had with animals as a little kid…* . In such cases, *animals as* would not be considered an N + P cluster, the preposition having an association with another preceding word, in this case the verb *had*.

In order to get an accurate picture of the various uses of *as*, each occurrence in the tagged version of the Corpus had to be checked individually. In almost all occurrences of prepositional *as* following nouns, the two words were separated from each other by some form of punctuation, either a period or a comma, further weakening the consideration of *as* as a nominal right colligate to nouns altogether. The most frequent use of *as* to post-modify a noun was in the phrase *such as* (383/3680). Thus, *as* was not considered for further analysis because its function as an adjectival complement to nouns on its own, as in *…cited their Christian faith as a reason…*, was quite limited and therefore irrelevant to a study focusing on frequent N + P clusters. The remaining occurrences of *as* in

constructions such as *as well (as)* and *as far as* were tagged as adverbials for 10% (446/3681) of the total tokens. In fact, the most common usage of *as* was in the double frame *as* + ADJ + *as* + NP + VP, with the first occurrence tagged as an adverb and the second tagged as a subordinating conjunction. As mentioned previously, *as* was removed from the top-ten list of prepositions from this study because it very rarely functions as a complement to nouns.

3.13 Nominal Left Colligates of *With*

Fewer than twenty words needed to be checked for nominal status in front of *with* (see Appendix F). *Deal with*, the most frequent two-word colligation including *with* as the right element occurred in four different forms, *deal with, deals with, dealing with, dealt with*, and almost always as verbals. The nominal left collocates of *with* occurring more than twenty times in the Corpus are:

45x *problem with*
42x *people with*
28x *relationship(s) with*
26x *touch with*
21x *war(s) with*

The collocation *touch with* represents part of the idiomatic prepositional phrase *in touch with*, so this item would be better thought of as a prepositional phrase following *keep, stay,* or *get* rather than as an N + P cluster.[20]

[20] For all left collocates of *with* occurring five or more times in the UGALECT Corpus, see Appendix F.

3.14 Nominal Left Colligates of *On*

Fewer than twenty words needed to be checked for nominal status in front of *on*. The nominal left colligates of *on* occurring more than twenty times in the Corpus are:[21]

53x	*war(s) on*	35x	*view(s) on*
50x	*effect(s) on*	24x	*information on*
47x	*impact on*	23x	*opinion(s) on*

3.15 Nominal Left Colligates of *By, From, At, About*[22]

In most cases, *by* functions as complement to a verbal participle or adjective such as is common in passive voice usage. There were very few nouns complemented with *by*. The most frequent N + P cluster with *by* was *article by* with 6 occurrences. Only one nominal left colligate with *from* occurred more than twenty times in the Corpus, *people from*, at 24 occurrences. Only 7 occurrences (9%) of *look at* were nominal, e.g. *take a look at*, in a total of 77 occurrences. So the most common, nominal left colligate with *at* occurred 12 times in the Corpus, *people at*. Only one nominal left colligate with *about* occurred more than 20 times in the UGALECT Corpus, *information about*, at 32 occurrences.

As can be seen in the above results, there are several very frequent prepositions in academic writing that serve as adjectival complements to nouns. Although *of* appears to be the most common preposition in N + P clusters, *to, in, for, with,* and *on* also have frequent nominal left colligates. The less frequent top-ten prepositions, *by, from, at,* and *about*, also occur less frequently as right colligates to nouns. In the next chapter, we will

---

[21] For all left collocates of *on* occurring five or more times in the UGALECT Corpus, see Appendix G.

[22] For all left collocates of *by, from, at,* and *about* occurring five or more times in the UGALECT Corpus, see Appendixes H-K.

take a closer look at the very frequent N + P clusters identified in the above analysis in order to determine the degree of attraction between certain nouns and certain prepositions in the UGALECT Corpus, thereby establishing the most robust of these two-word clusters.

CHAPTER 4

RESULTS & ANALYSIS

This chapter will focus on the results found for N + P clusters in the previous chapter and on the degree of attraction between certain high-frequency nouns and their prepositional right colligates as determined through a proportional analysis taking expectations of occurrence for particular prepositions into account. The research questions from Chapter 1 are addressed in turn as well.

4.1 Preposition and N + P Cluster Frequencies

The first research question was: What are the most frequent prepositions used by native speakers in freshman composition? The determination of the most frequent prepositions in the UGALECT Corpus was found through a raw word frequency count as generated by the concordancer (see Appendix A). The part-of-speech tags that were produced by the automatic tagger were also consulted in order to get an accurate picture of when certain words such as *to* were actually functioning as prepositions rather than as some other word class. For example, all occurrences of *to* functioning as an infinitive are labeled as such by the tagger, and the FIND command in Word can be used to search and count specific POS tags so that those words labeled and functioning as adverbials are not counted among the prepositions. In addition, certain concordances had to be checked manually through the concordancer for actual prepositional function.

The ten most frequent prepositions in the Corpus in descending order are *of, in, to, for, with, on, by, from, at,* and *about*. This finding is in line with expectations based on other studies and presentations of the most frequent English prepositions (Kennedy, 2003; Coffin & Hall, 1998; Francis, Kučera & Mackie, 1982).

The second research question was: What are the most frequent nominal left colligates of the ten most frequent prepositions in freshman composition, and what are the frequencies of occurrence of these two-word phrases (N + P clusters) in the Corpus of freshman essays? This research question was answered by using the cluster function in the concordancer to rank the frequencies of each preposition as the right collocate in any two-word sequences in the Corpus (see Appendices B - K). Some individual two-word sequences also had to be checked manually, such as *work in*, in order to determine nominal, verbal, or adjectival functions of those left collocates. In sorting the concordance lines alphabetically by the immediate left collocates for each preposition, the co-text reveals the actual part of speech of each left collocate in use.

The third research question was: Are these nouns usually followed by prepositions in the Corpus, and, if so, which prepositions are their most frequent right colligates? In other words, what proportion of these nouns is post-modified by a particular preposition as opposed to some other word class or some other preposition? By targeting the nouns found in the previous step through the concordancer, all right collocates of these nouns could be sorted alphabetically and proportions of prepositions as immediate right colligates could be determined. Also, by using the N-gram function in the concordancer, which can be used to rank all two-word frequencies in the Corpus, the raw frequencies for all two-word sequences could be verified (see Appendix L).

4.2 N-Grams and Proportional Analysis

The concordancer has an N-gram function which allows for frequency counts of words in a contiguous sequence (phrases) in a corpus without regard to grammatical structure. An N-gram search yields the most frequent 2-word, 3-word, 4-word, 5-word, and so forth sequences. If set at 2-word sequences, the N-gram function in a concordancer lists and counts every 2-word sequence in a corpus, and every word is part of a 2-word sequence exactly twice and part of a 3-word sequence thrice and so on. For example, the phrase *the fact of the matter is* will yield the 2-word sequences *the fact, fact of, of the, the matter,* and *matter is*. The concordancer tracks the frequency of occurrence for each sequence and then lists them in rank order from most frequent to least frequent. The most frequent two-word sequence in the UGALECT Corpus is *of the*. The N-gram function was used to rank all two-word N + P clusters occurring ten times or more in the UGALECT Corpus (see Appendix L). A proportion test was then established by assigning expected frequencies of occurrence for each of the top-ten prepositions below.

A t-score (Stubbs, 2002) is a simple measure of whether a particular rate of occurrence is in line with expectations or not. When a lexical sequence occurs at a greater than expected rate, that sequence is considered statistically significant. First, the actual rate of occurrence must be established; then using a basic formula of probability, the expected rate is calculated and compared to the actual rate. For example, the frequency of the noun *part(s)* alone is 328, and the frequency of *of* alone is 14,742, and the frequency of *part(s) of* as a sequence is 247 in the UGALECT Corpus. So at any given point in the Corpus, the probability of either *part* or *parts* being the next word is 328/500,000 = .000656 (about .07%), and the probability of *of* being the next word is much greater at 14,742/500,000 = .0295 (about 3%). So the probability of the two words

occurring together in either order is .000656 x .0295 = .000019352 (about .002%). And the probability of them occurring in the sequence *part(s) of* is half that: .000019352/2 = .000009676 (about .001%). The actual, observed frequency of *part(s) of* in the Corpus is 247/500,000 = .000494 (about .05%). So, the observed frequency is 50 times greater than what would be expected by chance (.000494/.000009676). This is certainly a significant rate of occurrence for this two-word sequence. The distribution of words in a text is not random, however.

This method does not take into consideration whether the occurrences found are in line with expectations for each of the top-ten prepositions in relation to each other. In other words, we should first establish an expected rate of occurrence for each preposition based on their actual rate of occurrence as opposed to the actual rate of occurrence of the other top-ten prepositions. A proportional analysis using expected frequency ratios (input probabilities) for each of the top-ten prepositions sets the bar a bit higher in determining the most robust colligations. Because *of* is a very common word in English, its occurrence as a very frequent nominal colligate is not surprising. Therefore, expectations for the occurrence of *of* in any environment should be considered based on its relative frequency with regard to the other most frequent prepositions that could go in its place. Only insofar as *of* is found in much higher numbers than what is to be expected from its relative frequency ratio should its collocations be regarded as significant and worthy of our attention.

One way to determine whether the frequent N + P clusters found above warrant attention in ESL/EFL writing classrooms is to do proportion tests in order to see what percentage of a noun's occurrence is actually followed by a particular preposition as opposed to any other frequent preposition. If the distribution of words in a language were

completely random, we could generate an expectation of occurrence for any word based on its actual frequency in a given corpus. In order to do this, a percentage of expected frequencies for the top-ten prepositions in the corpus was set up as follows: the number of occurrences of each word tagged as a preposition by the automatic tagger was recorded, and the total of those occurrences was used as a factor in determining a relative expected frequency of occurrence (input probability) for each preposition as compared to the other top-ten prepositions in Table 4.1:

Table 4.1 Input Probability for the Top-Ten Prepositions[23]

PREPOSITION	OCCURENCES	PERCENTAGE
of	14,742	31%
in	9399	20%
to	6251	13%
for	4395	9%
with	3459	7%
on	3282	7%
by	2005	4%
from	1871	4%
at	1307	3%
about	1134	2%
TOTAL	47,845	100%

This total demonstrates that just these top-ten prepositions make up almost 10% (47,845/500K) of the entire UGALECT Corpus, which is in line with expectations given the frequency of this word class in the formal, written register of around 11-13% for all prepositions (Biber et al., 1999; Kennedy, 2003) So, if the distribution of words in the corpus were completely random, we would expect the most frequent preposition, *of*, to

[23] This list includes all prepositions occurring 1000 times or more in the UGALECT Corpus, except for *as*, which functions much more frequently as a subordinator or as a correlative adverbial, i.e. *as + ADJ + as + NP*, rather than as a nominal post-modifier, eliminating it from consideration as a frequent N + P cluster component.

show up a little over 30% of the time compared to any of these other top-ten prepositions. Therefore, each preposition's occurrence will now be judged in relation to its established rate of occurrence in Table 4.1 above.

By looking at the immediate right collocates of the nouns suspected of being phrasal from the frequencies determined in the last chapter, we can discern whether the occurrence of a particular noun with a particular preposition is in line with, or greater than or less than, what can be expected from the above percentages. Only those prepositions occurring with a much greater than expected ratio as immediate right colligates to high frequency nouns were then considered robust N + P clusters.

For example, by looking at the concordance lines of the most frequent N + P cluster in the Corpus, *part(s) of,* we can see that the lemma[24] PART (either as *part* or as *parts*) occurs 328 times and with *of* as its immediate right collocate 235 times. So, *of* is the right colligate for *part(s)* in 235 out of 328 total occurrences or 72% of the time. This percentage is more than twice as much as would be expected from the ratio of occurrences of *of* in the chart above (31%). The lemma PART occurs followed by some other preposition in the top-ten list only 10% of the time (34/328), the most frequent of which is *in*; in a total of twenty-five tokens, 9 were TAKE *part in* and 9 were PLAY *a part in*. And the lemma PART occurs followed by something other than one of the top-ten prepositions above 18% of the time (59/328). So, there is a great amount of attraction between PART and *of*, a robust finding, which supports regarding it as a single lexical unit. In other words, the occurrence of *part(s)* says something about the occurrence of *of*

[24] A lemma is an abstract category of all the forms of a word; in this case, it includes all singular and plural forms of the noun, PART.

in that we can generally expect *of* to occur in the wake of this particular lemma in a much greater than expected proportion when it is not part of a multi-word verb with *in* as noted above.

On the other hand, another very frequent noun in the Corpus, *way(s)*,[25] occurs as a left colligate to all of the top-ten prepositions in the Corpus, with the greatest numbers in the following ratios in Table 4.2:

Table 4.2 Prepositional Right Colligates of *way*

WAY	**RAW FREQUENCY**	**PERCENTAGE**
	1001	100%
way(s) of	107	11%
way(s)in	33	3%
way(s) to	22	2%
way(s) for	18	2%

All of these percentages are lower than what would be expected in a random distribution of each of these prepositions. Obviously, the occurrence of *way(s)* does not indicate the occurrence of any particular preposition in its wake.

Another consideration is for highly topical nouns such as *war*, which has a high rate of occurrence in the Corpus, but usually occurs in the timely collocations *war in Iraq* and *war on terror*. Also, *weapons* is found most frequently in *weapons of mass destruction*. Such nouns as *war(s), weapon(s), candidate(s),* and *debate(s)* are particularly frequent in this particular Corpus because of the fact that these essays were written during the Iraq War and in a presidential election year, just as the proper nouns *Obama* (285 tokens), *Clinton* (117 tokens), and *McCain* (36 tokens) are indicative of such 'situated discourse.'

---

[25] For a closer look at the behavior of the very frequent noun *way* and its collocates, see Sinclair (1999).

This type of relative analysis was conducted on all frequently occurring N + P clusters (those occurring twenty times or more), which are listed according to absolute frequency in Appendix L as two-word clusters or N-grams. Those having a frequency ratio of double the input probability with a particular top-ten preposition are noted as warranting consideration as extremely robust N + P two-word clusters going forward. Furthermore, those robust two-word clusters also had to be checked for status as frequent three-word clusters, so a cutoff of 75% of two-word clusters occurring as three-word clusters was also applied. For example, if a two-word cluster such as *addition to* occurs in over 75% of its occurrences as *in addition to*, which it does, it was eliminated from further consideration as a pure N + P cluster.

The N + P clusters with *of* occurring 20 times or more along with their frequency ratios are:

OF: Input Probability = 31%

thousands of	88% (64/73)	*weapons of*	51% (68/134)
amount(s) of	87% (180/207)	*development of*	51% (31/66)
variety/ies of	86% (37/43)	*advantage(s) of*	51% (21/51)
majority of	85% (76/89)	*pound(s) of*	49% (22/45)
type(s) of	83% (224/269)	*form(s) of*	48% (130/269)
millions of	83% (52/63)	*understanding of*	48% (30/63)
declaration of	83% (20/24)	*source(s) of*	45% (44/98)
lack of	80% (82/103)	*means of*	45% (27/60)
sort(s) of	79% (45/57)	*principle(s) of*	45% (20/44)
kind(s) of	78% (108/138)	*result(s) of*	44% (77/174)
part(s) of	73% (247/340)	*risk(s) of*	43% (35/82)
aspect(s) of	67% (136/202)	*member(s) of*	43% (34/79)
front of	64% (43/67)	*beginning(s) of*	41% (45/109)
number(s) of	63% (140/223)	*production of*	41% (22/54)
lot(s) of	61% (128/210)	*couple of*	41% (20/49)
definition of	61% (23/38)	*method(s) of*	40% (35/87)
importance of	57% (45/79)	*cause(s) of*	38% (36/95)
loss of	57% (29/51)	*piece(s) of*	37% (45/122)
creation of	57% (26/46)	*fear(s) of*	36% (26/73)
sense of	54% (79/146)	*top(s) of*	34% (33/97)
percent of	54% (60/111)	*danger(s) of*	34% (22/64)

group(s) of	33%	(74/227)	*picture(s) of*	17%	(45/261)
half of	33%	(26/78)	*case(s) of*	17%	(43/251)
cost(s) of	33%	(23/69)	*freedom(s) of*	17%	(25/147)
chance(s) of	32%	(38/120)	*content(s) of*	17%	(21/123)
period(s) of	32%	(36/111)	*point(s) of*	16%	(46/280)
quality/ies of	32%	(32/99)	*view(s) of*	16%	(43/272)
appearance(s) of	31%	(35/112)	*middle of*	16%	(23/147)
hope(s) of	31%	(26/84)	*knowledge of*	16%	(22/141)
list(s) of	31%	(25/80)	*role(s) of*	15%	(27/176)
rest of	30%	(58/196)	*area(s) of*	15%	(23/152)
example(s) of	29%	(100/340)	*future of*	14%	(24/173)
benefits of	29%	(24/83)	*color(s) of*	14%	(23/162)
state(s) of	28%	(91/323)	*name(s) of*	14%	(21/155)
effect(s) of	28%	(57/207)	*story/ies of*	13%	(32/240)
purpose(s) of	28%	(55/194)	*citizen(s) of*	13%	(30/227)
side(s) of	28%	(48/169)	*president of*	13%	(28/222)
meaning(s) of	28%	(28/101)	*opinion(s) of*	13%	(23/182)
course of	28%	(27/97)	*nature of*	13%	(21/168)
center(s) of	27%	(27/100)	*way(s) of*	11%	(107/1001)
idea(s) of	26%	(116/440)	*end of*	11%	(65/607)
style(s) of	26%	(39/152)	*issue(s) of*	11%	(43/394)
age(s) of	25%	(64/252)	*death(s) of*	11%	(33/309)
hours of	25%	(26/105)	*attention of*	10%	(23/230)
use(s) of	24%	(185/761)	*life/lives of*	9%	(99/1135)
level(s) of	24%	(32/134)	*word(s) of*	8%	(26/334)
population of	24%	(23/95)	*message(s) of*	8%	(20/254)
set(s) of	24%	(20/83)	*year(s) of*	7%	(49/727)
history of	23%	(45/192)	*act(s) of*	6%	(34/593)
control(s) of	23%	(32/137)	*problem(s) of*	6%	(31/516)
term(s) of	23%	(30/133)	*day(s) of*	6%	(27/478)
sign(s) of	23%	(21/90)	*time(s) of*	5%	(58/1198)
goal(s) of	22%	(27/125)	*people of*	4%	(90/2344)
feeling(s) of	21%	(29/140)	*world of*	4%	(34/911)
evidence of	21%	(20/97)	*right(s) of*	4%	(20/500)
leader(s) of	19%	(20/108)	*debate(s) of*	3%	(23/786)
image(s) of	18%	(60/329)	*life/lives of*	2%	(28/1135)
generation(s) of	18%	(25/139)	*process of*	2%	(25/125)
need(s) of	18%	(23/130)	*women/woman of*	2%	(24/1282)
city/ies of	18%	(21/120)	*man/men of*	2%	(21/991)

All of the N + P clusters above occurred over twenty times in the Corpus. The cutoff for the most robust N + P clusters was set at twice the input probability, which for *of* is 31%. So, only those N + P clusters having double this input probability (62%) or above for

their frequency ratios will be considered extremely robust and will be taken up again in the semantic taxonomy below.

The high-frequency N + P clusters with *in* along with their respective frequency ratios are:

IN: Input Probability = 20%

increase(s) in	69%	(54/78)	*problem(s) in*	6%	(32/516)
role(s) in	35%	(61/176)	*technology/ies in*	6%	(20/328)
change(s) in	33%	(121/367)	*people in*	4%	(99/2344)
interest(s) in	30%	(34/113)	*children in*	4%	(36/1024)
difference(s) in	22%	(32/144)	*student(s) in*	4%	(27/646)
war(s) in	14%	(110/791)	*thing(s) in*	4%	(27/675)
point(s) in	12%	(33/280)	*country/ies in*	4%	(24/577)
character(s) in	12%	(20/171)	*day(s) in*	4%	(20/478)
place(s) in	10%	(36/366)	*life/live(s) in*	3%	(34/1135)
situation(s) in	10%	(22/225)	*way(s) in*	3%	(33/1001)
part(s) in	8%	(25/32)	*time(s) in*	3%	(40/1198)
issue(s) in	7%	(27/393)	*school(s) in*	3%	(22/780)
women/woman in	6%	(75/1282)	*debate(s) in*	3%	(20/786)

Only *increase(s) in* occurs with more than double the input probability for *in* (40%). Although not one of the top-ten prepositions under analysis here, *between* deserves honorable mention because it occurs as the right colligate to *difference(s)* for almost one-third of this noun's total occurrences: *difference(s) between* = 28% (40/144), which is more significant than *difference(s) in* (22%) above, although with a much less frequent preposition. In fact, Kennedy (1991) found *difference* to be the most frequent left collocate of *between* in the Lancaster-Oslo/Bergen (LOB) one-million-word corpus of written British English.

The high-frequency N + P clusters with prepositional *to* in the Corpus along with their respective frequency ratios are:

TO: Input Probability = 13%

access to	53%	(39/73)	*solution(s) to*	38%	(39/102)
addition to	51%	(26/51)	*answer(s) to*	21%	(23/107)
response(s) to	40%	(27/68)	*attention to*	13%	(30/230)

Four of these frequent N + P clusters demonstrate higher than double the input probability for prepositional *to* of 26%. Although *addition to* and *response(s) to* are very robust colligations, they are used very frequently as *in addition to* (20/26 = 77%) and *in response to* (20/27 = 74%), which would be better thought of as one of several common three-word prepositionals having the pattern *in* + N + P (see Appendix L). With a cutoff of 75% of occurrences in longer three-word clusters having been set, this eliminates *addition to* but retains *response(s) to* as a very robust N + P two-word cluster. *Answer(s) to* is not quite high enough to make the cutoff for *to*, but it occurs quite frequently as *answer(s) to (the) question(s)* for 44% of the *answer(s) to* occurrences, which may be a pedagogically useful collocation. *Attention to* is in line with expectations for prepositional *to*, but *attention* is also post-modified by several other top-ten prepositions in the Corpus, including *of, on, in,* and *from*, so considering *attention to* a lexical unit in and of itself would not be warranted and could even be confounding for students. Focusing on the preceding verbs, the extended collocations include GET/KEEP *the attention of someone* and FOCUS *attention on*. Those collocations of note with *to* here are the verb forms PAY *attention to* (43%), DRAW *attention to* (30%), and BRING *attention to* (10%) for pedagogical consideration.

The high-frequency N + P clusters with *for* are:

FOR: Input Probability = 9%

room for	30%	(23/76)	*order for*	9%	(29/327)
need for	27%	(35/130)	*candidate(s) for*	5%	(23/510)
reason(s) for	22%	(77/355)	*time for*	3%	(28/964)
plan(s) for	12%	(23/189)	*life/lives for*	2%	(28/1135)

Three nouns demonstrate a robust attraction with *for* at more than double the input probability (18%) and will be taken up again below. Half of the occurrences of *plan(s) for* as a noun occur as either *Obama's, Barack's,* or *his* (referring back to Barack Obama) *plan(s) for*, and half of the occurrences of *candidate(s) for* are followed by *president* or *presidency*. *Order for* always occurs as part of the common three-word prepositional pattern *in* + N + P (see Appendix L for others).

The high-frequency N + P clusters with *with* are:

WITH: Input Probability = 7%

relationship(s) with	13%	(28/218)	*war(s) with*	3%	(21/791)
problem(s) with	9%	(45/516)	*people with*	2%	(42/2344)

None of these nouns demonstrates a frequency ratio with *with* at double its input probability (14%), and *problem(s)* also occurs followed by *of* and *in*, although at lower frequencies and at lower input probabilities than expected for those respective prepositions. This is a case in which it would be better to consider three N + P clusters together: *problem(s) with, problem(s) of,* and *problem(s) in*, along with a discussion of their respective distributions of use, in other words, their repeated concordances such as *the problem with this is that*, *the problem of illegal immigration,* and *a problem in the United States*.

The high-frequency N + P clusters with *on* are:

ON: Input Probability = 7%

impact(s) on	55%	(37/67)	*view(s) on*	12%	(35/285)
effect(s) on	24%	(50/207)	*war(s) on*	7%	(53/791
opinion(s) on	13%	(23/182)	*information on*	6%	(24/377

On is very robust as a prepositional right colligate to *impact(s)* and *effect(s)*. However, *effect(s)* also occurs with a high frequency followed by *of*. Although *effect(s) of* occurs slightly more frequently (57/207 tokens = 28%) than *effect(s) on* (50/207 tokens = 24%) in the Corpus, the former was not over double the input probability set previously for *of* (62%), and the latter was above double the input frequency for *on* (14%). *Opinion(s) of* occurs exactly the same number of times as *opinion(s) on* in the Corpus indicating a need for more context to clarify their respective distributions of use. *View(s) of* occurs more frequently than *views on*, but both were lower than double the input probability for their respective prepositions. These are all examples of pairs of N + P clusters that would require greater contextualization with extended concordances and a greater focus on distinctions in their respective ranges of use: *impact(s) of/on, effect(s) of/on*, *opinion(s) of/on*, and *view(s) of/on*. In other words, further consideration of the differences in their patterning along the vertical dimension (the paradigmatic axis) in the respective concordance lines (the syntagmatic axis) is warranted for pedagogical applications.

The only high-frequency N + P cluster with *by,* which has an input probability of 4% in the Corpus, is *article(s) by* at 4% (6/169). *By* here, of course, means *written by* and does not occur as a nominal right colligate at a significant rate. This preposition proves to be much more useful as a right colligate to passive verbs and adjectives (see Appendix H) and should therefore continue to be taught as such in academic contexts.

The only high-frequency N + P cluster with *from*, which has an input probability of 4%, is *people from* at 1% (24/2344). *People* is a noun that is very frequent in the

Corpus (2344 tokens) and frequently followed by many different prepositions, yielding no strong colligations with any of the top-ten prepositions.

There were no high-frequency N + P clusters with *at*, and this preposition proves not to be very useful as a nominal right colligate. See Appendix J for some N + P clusters with *at* that did not meet the cutoff rate for frequency in this study such as *issue(s) at (hand)* and *chance at (winning)*.

The only high-frequency N + P cluster with *about*, which has an input probability of 2%, is *information about* at 9% (33/377), a robust cluster. *About* shows a very significant attraction with the noun *information* at more than four times the input probability for *about*, but *information on* from above was also quite frequent though not so robust for that particular preposition. *Knowledge* warrants mentioning as well because of its high rate of occurrence followed by *about*, 6% (8/141). This preposition shows significant attraction with the noun *knowledge* at three times its input probability, although *knowledge* itself is a much less frequent noun than *information* in the Corpus.

The most robust N + P clusters from above would be good candidates for inclusion in the second language writing curriculum. ESL/EFL writing students are explicitly taught how to use transitional expressions such as *for example* and *on the other hand* in their academic writing classes because these are frequent and useful transitional devices for, especially written, academic discourse (for example, see Oshima & Hogue, 2006). Given the high relative frequencies of N + P clusters in academic writing in general, these prepositions could also be thought of as cohesive devices among nouns and their adjectival post-modifiers. For instance, *as a result* (50 tokens) and *result(s) of* (77 tokens) account for 73% of all occurrences of nominal *result(s)* (174 tokens) in the

Corpus. These two most frequent environments for the noun *result(s)* could be presented to L2 writers together with explanations as to their respective distributions of use.

4.3 Qualitative Analysis: A Semantic Taxonomy for N + P Clusters

Next, we will look at a semantic taxonomy of N + P clusters as a way to facilitate the presentation, learning, and retention of these common structures for the benefit of non-native speakers based on extant presentations of multi-word verbs in ESL/EFL textbooks and collocational frameworks in the relevant research. Traditional presentations would include organizing the vocabulary alphabetically by noun or by preposition, semantically by relevant meaningful contexts, or by frequency. As noted by both Sinclair (1991b) and Lewis (2000), *of* is by far the most frequent prepositional right colligate to many common nouns. And many of these N + P clusters can be grouped under one functional heading, that of 'quantifiers,' what Morenberg (2002) calls 'prearticles' such as *lot(s) of*. The frequent N + P clusters with *of* above that proved most robust because of a higher than expected ratio of occurrence with *of* are: [26]

part(s) of	*sense of*
type(s) of	*majority of*
amount(s) of	*thousands of*
number(s) of	*millions of*
aspect(s) of	*sort(s) of*
kind(s) of	*variety/ies of*
lack of	

[26] The N + P cluster *declaration of* was eliminated from further consideration because 85% of its occurrences were as part of the proper noun *Declaration of Independence*; also, *front of* was eliminated because it was realized as *in front of* in 95% (all but 2) of its occurrences; *top of* was eliminated because all of its occurrences were in *on top of*; and *advantage(s) of* was eliminated because 76% of its occurrences were as part of the phrasal verb *TAKE advantage of*.

In revisiting Cortes' findings (2002) with regard to four-word lexical bundles in freshman composition (see Section 2.4), we find the following N + P clusters from above: *lot of, part of, variety of, result of,* and *form of*, so these N + P clusters are particularly robust in first-year composition writing.

The fourth research question formulated for this study was to determine whether or not other frequent prepositions would prove to be as robust as *of* as nominal right colligates in the written academic register of native speakers. Although there are some strong colligations with other top-ten prepositions such as *in, to, about, for,* and *on*, *of* proved to be the most frequent, robust, and highly distributed nominal right colligate by far.

The rest of the top-ten prepositions and their most frequent nominal right colligates occurring at higher than expected ratios are:[27]

increase(s) in	*room for*
solution(s) to	*effect(s) on*
access to	*impact on*
reason(s) for	*information about*
need for	

The following semantic taxonomy is based on a previous *ad hoc* classification of nouns by Butler (1998), who was looking at collocational frameworks for nouns in Spanish speech, transcribed interviews, and newspaper articles, the latter being his one written corpus from Spain's national daily, *El País*. The focus of Butler's study was on nouns that occur in the frames *un/una/el/la* _____ *a/de/en/por* in five different corpora. The twelve semantic categories used in his study for nouns occurring in front of *de*

---

[27] The phrase *addition to* was eliminated from further consideration as an N + P cluster because 77% of its occurrences were as part of the three-word prepositional pattern *in* + N + P (*in addition to*), a frequent prepositional phrase used as a cohesive device in academic writing.

(English *of/from*) are: measure/quantity, kind/manner, place, time, process/plan, matter, part/stage, organization, sense, existence, human, and modal concepts. In the spoken corpora, Butler found a prevalence of "nouns referring to important features of everyday life (*sala, puerta, plaza, iglesia, calle, universidad,* etc.)" (p.17), whereas in his written corpus, he found an abundance of abstract nouns, which he described as "characteristic of formal written style…not found to any extent in spoken Spanish" (p.18).

The five semantic categories below were adapted from Butler's model to classify the most robust N + P clusters from the UGALECT Corpus based on what little meaning they carry out of context: quantity/measure for amounts, quality/kind for general classifications, matter/sense/knowledge for mental concepts, process/plan for causal or procedural relationships, and modal concepts for possibilities or necessities.[28]

All of the N + P clusters in the quantity/measure group express amount or numeric determinations for their following objects:

- Quantity/Measure: *thousands of, amount(s) of, majority of, millions of, lack of, number(s) of,* and *increase(s) in*.

The quality/kind N + P clusters express some form of grouping or general way to classify their following objects:

- Quality/Kind: *variety/ies of, type(s) of, sort(s) of, kind(s) of, part(s) of, aspect(s) of,* and *kind(s) of*.

The matter/sense/knowledge category includes all N + P clusters expressing mental conceptions of their following objects:

---

[28] For very finely detailed semantic groupings of N + P structures organized by preposition, see Francis, Hunston, & Manning (1998).

- Matter/Sense/Knowledge: *solution(s) to* and *information about*.

The N + P clusters in the process/plan group express some form of causal or procedural relationship with their following objects:

- Process/Plan: *response(s) to, effect(s) on, reason(s) for*, *plan(s) for,* and *impact on*.

The modal N + P clusters express possibilities, probabilities, obligations, or necessities:

- Modal Concepts: *access to, need for* and *room for*.

Interestingly, there were no occurrences of robust N + P clusters that were used to refer to people or humans, time, place, organizations, nor part/stage (Butler's other categories), which indicates that such topics may be more common in speech.

Such a semantic/functional taxonomy also represents one way of categorizing frequent N + P clusters for ESL/EFL writers. For example, when writing process or cause/effect essays, students could be given some exposure to the N + P clusters in the process/plan grouping above, and when writing classification or comparison/contrast essays, some exposure to kind/quality N + P clusters would be beneficial; when writing argumentative essays, those in the matter/sense/knowledge group would be useful. N + P clusters in the quantity and quality groups are the most common and the most generally applicable to differing contexts, and the learners in this study have demonstrated some familiarity with these forms.

4.4 Learner Usage of Frequent and Robust N + P Clusters

In answer to the fifth research question regarding the occurrence of robust N + P clusters from above in the NNS essays, some of these N + P clusters were found to occur

at least once in the learners' writing. The robust N + P clusters that do occur in the learner data along with their number of occurrences in descending order of frequency are:

N + P Clusters in the Learner Essays

kind(s) of	35x	*majority of*	3x
part(s) of	21x	*solution to*	3x
type(s) of	14x	*lack of*	2x
amount of	6x	*number of*	2x
sort(s) of	4x		

The robust N + P clusters that occur in the learner data only once (*hapax legomena*) are: *aspect of*, *millions of*, *thousands of*, and *variety of*. The robust N + P clusters used by none of the L2 students are: *increase(s) in, access to, impact(s) on, reason for, room for, need for,* and *information about*.

By mapping the semantic categories of the robust N + P clusters from above onto those used and not used by the non-native speakers, we can further analyze the learner usage. It is interesting to note here that the learners are using N + P clusters in the greatest numbers from the quantity and quality semantic categories above, the most numerous and perhaps the most concrete semantic categories, whereas those N + P clusters that are less abundant in the learner essays are found in the semantic categories of matter/sense/knowledge and process/plan, the more abstract categories. Also, learners made no use of the robust N + P clusters from the modal category.

In revisiting the actual learner errors with regard to preposition use after nouns from Section 3.2 and comparing them to frequent N + P clusters, Spanish speakers used **problem of everyone, *reasons of it, *opinion in the situation, *importance in control themselves, *interest about, *decrease on,* and **effects to*. These are all nouns that occurred at extremely high frequencies in the NS Corpus with an appropriate preposition: *problem(s) with, reason(s) for, opinion(s) on, importance of, interest(s) in, decrease(s) in,*

and *effect(s) on*. Only this last N + P cluster with the appropriate preposition occurred in the Spanish-speaker data, but this is a very limited data set (approximately 22K words) compared to the NS Corpus.

Errors from the speakers of Asian languages with regard to prepositions after nouns occurring in high frequencies in the NS Corpus were **reasons on that*, **a big role of music*, **a period time*, **thousands miles away*, **attention on the children* and **understanding to freedom*. Again, these are all nouns that occurred at extremely high frequencies in the NS Corpus with an appropriate preposition: *reason(s) for, role(s) in, period(s) of, thousands of, attention to,* and *understanding of*.

With regard to the various academic word lists discussed in Section 2.1, the only two nouns from robust N + P clusters not appearing on any of the lists (the GSL, UWL, and AWL) were *thousands* and *millions*. Nouns in robust N + P clusters on the UWL and AWL, which are very similar lists consisting of higher level academic headwords, were *aspect(s)*, *access*, and *impact(s)*. The remainder of the nouns from the most robust N + P clusters all appear on the General Service List (GSL), which is a list of the 2000 most common words in a 5-million word academic corpus. It is also worth noting here that all of the singular forms of the nouns in the most robust N + P clusters occurred within the first 1500 words of the UGALECT Corpus when ranked by raw frequency.

4.5 Nominal Density and Preposition Density

The sixth research question asks about the preposition and nominal densities[29] of the NS academic writing in comparison to that of the NNSs. This question is posed because we already know that prepositions are one of the most frequent word classes in English, that learners occasionally omit them, and that nouns are a relatively frequent word class in formal, academic writing as compared to conversation (see Section 2.4). Hudson (1994) found common nouns to represent 24% of the Brown Corpus of one-million words of written American English.

In calculating the nominal density of the NS Corpus, the number of common nouns as tagged by the part-of-speech tagger was divided by the total word count. The NS essays showed a nominal density of approximately 23% (114,075/500K) and a preposition density of over 11% (57,241/500K), which is in-line with expectations for these word classes in the formal, written register (see Sections 2.3-2.4).

The Spanish-speaker essays had a nominal density of about 20% (4208/21,483) and a preposition density of 9.6% (2063/21,483). So, the Spanish speakers in this study used a lower percentage of both nouns and prepositions than did the native speakers. Also, their preposition selection proved to be a problem on occasion (see Section 3.2). Both the preposition and nominal densities in the Spanish-speaker essays are not as high as they could be for academic writing. Of course, this could be due to the fact that the Spanish speakers had a limited time frame in which to produce their essays, and Moreno

[29] For nominal density, we are only considering common nouns here such as those found in the most frequent N + P clusters, whether singular or plural.

(2008) cautions against making any strict comparisons between two corpora having more than one feature in contrast.

The speakers of Asian languages produced text with a nominal density of 23% (3199/13,727) and a preposition density of a little over 9.5% (1309/13,727). This group of NNSs was found to be using common nouns at a rate in-line with that of native speakers. In her quantitative study of the academic writing of college students, both native and non-native speakers, Reid (1988) also found that Chinese college students were using nouns at the same rate as the NSs in her study. However, the rate of usage of prepositions by the Asian students was found to be below that of NSs in both Reid's and the present study.

Both groups of learners are using prepositions at a lower rate than native speakers, and the Spanish speakers are using common nouns at a lower rate than native speakers. It is suggested here that some attention to N + P clusters in the second language writing curriculum could address two deficiencies at once, that of preposition density and selection and that of nominal density or content vocabulary.

CHAPTER 5

SUMMARY & IMPLICATIONS

5.1 Summary

In Chapter 1, we laid out the foundations for the current focus on N + P clusters by analogy to multi-word verbs and adjectives, which are already covered in extant ESL/EFL grammar and vocabulary textbooks and various reference manuals (see Section 1.1). It was proposed that an awareness of N + P clusters on the part of ESL/EFL students could help alleviate the burden students have in two areas of sentence construction, preposition selection in English and nominal density in academic writing. A collocational approach to prepositions that follow and modify nouns was outlined as a way to present prepositions in their most frequent lexico-grammatical environments, thereby making them more salient to learners.

Chapter 2 presents a brief overview of the history of academic word lists for NNSs (see Section 2.1) and reviews the literature on ESL errors in academic writing with regard to prepositions, which rank very highly among error type frequencies even for advanced learners (see Section 2.2). Chapter 2 also describes the various corpus studies on native-speaker English usage and lexical bundles in academic discourse in particular, which evidence an abundance of N + P clusters (see Sections 2.3-2.4).

In Chapter 3, the use of prepositions as complements to noun phrases was analyzed both in L1 and L2 academic writing. Primary evidence of preposition errors was presented. In the qualitative analysis of Spanish-speaker academic writing, errors with prepositions were found in all environments, i.e. in prepositional phrases and after

verbs, adjectives, and nouns, and such errors were manifested in various ways, e.g. as errors in the selection, epenthesis, and omission of prepositions (see Section 3.2). Speakers of various Asian languages were also found to be misusing English prepositions after nouns in their academic writing with the most frequent error type being omission of the requisite English preposition altogether.

An extensive quantitative extraction of two-word sequences in the form of N + P clusters as used by NSs in their academic writing was the primary focus of this research in Chapter 3 because nouns and prepositions have been shown to be especially dense in the register of academic writing (see Section 2.4), and it is felt that NNSs could benefit from some focus on N + P clusters in a lexical syllabus for college-level writing. First, raw preposition frequencies were established and then their most frequent nominal left colligates were isolated from the UGALECT Corpus (see Sections 3.6-3.16).

Prepositional phrases, i.e. P + NP structures such as *on the other hand* and *in fact*, functioning as transition signals and conjunctive adverbs in academic writing are presented in extant ESL/EFL teaching materials (for example, see Oshima & Hogue, 2003, pp. 295-299). Yet N + P clusters also represent robust lexical units, as demonstrated by the strong attraction found between certain frequent nouns and prepositions in the UGALECT Corpus of NS freshman essays (see Section 4.1 & 4.2). Also, Gitsaki (1999) found N + P structures easier for learners to grasp and retain than P + NP structures, perhaps because of the greater salience of content words such as nouns being encountered first in the syntagmatic sequence.

Some potential N + P clusters such as *addition to*, *attention to, advantage of,* and *declaration of* were eliminated from further consideration as such because they were found to be functioning as parts of greater phraseological units such as prepositional

phrases, extended verb phrases, or proper noun phrases most of the time: *in addition to*, PAY/GIVE *attention to*, TAKE *advantage of,* and *Declaration of Independence*. In the qualitative analysis, the most frequent and robust N + P clusters that occurred with much higher than expected ratios for each preposition were then grouped into a semantic taxonomy as one way to present them in L2 writing classes with relevance to their potential for use in particular essay types (see Section 4.3). Because the written academic register is marked by a preponderance of N + P clusters, NNSs would be well-served to have their attention drawn to these structures both in their academic reading and in the form of phrase lists such as those provided for multi-word verbs and adjectives in pedagogical materials.

N + P clusters as lexico-grammatical units are more indicative of the formal, written register than of conversation; this has been repeatedly verified by Biber (1988, 2006), Biber and Clark (2000), Biber and Conrad (1999), Biber, Conrad, and Cortes (2003, 2004), Biber et al. (1999, 2002), and other independent researchers such as Coxhead and Byrd (2007), Reid and Byrd (1998), Halliday (1991), Kennedy (2003), Sinclair (1991b), and Sinclair and Carter (2004).

The NNSs in this study were also found to be using some robust N + P clusters in their academic writing lending further credence to their treatment as lexical units; however, the learners demonstrated their ability for using N + P clusters in the semantic categories of quantity and quality such as *amount(s) of, increase(s) in, part(s) of, kind(s) of,* and *type(s) of* to a greater extent than N + P clusters in other semantic categories such as modal concepts (*access to*) and the plan/process group (*effect(s) on*), perhaps the more abstract categories in need of greater contextualization. Specific learner errors were also found to be made in certain robust N + P clusters as used commonly by native speakers

(see Section 4.4). Thus, although the learner data was scant in comparison to the NS Corpus, these learners demonstrated a lack of awareness of the usage conventions of particular prepositions with certain very high-frequency nouns in formal, written English.

We find as well essential differences in the types of errors non-native speakers make in their academic writing and those of native speakers as found in research on error types in academic writing (see Section 2.2). There are "a number of features which point to systematic lexico-grammatical differences between native-speaker English and ELF, for example omitting definite and indefinite articles, insertion of prepositions (e.g. *can we discuss about this issue*)" (O'Keefe, McCarthy, & Carter, 2007, p. 28), and omission or inaccurate selection of English prepositions, as we saw in the primary evidence for this study (see Section 3.2). Certain types of lexico-grammatical errors are limited to NNSs, i.e. native speakers just do not tend to make such errors. Function words like articles, prepositions, and conjunctions are particularly challenging for adult learners, while they are largely selected subconsciously by native speakers, who would be hard-pressed to come up with any hard and fast rules with regard to their own usage. Furthermore, research has shown that collocations and multi-word units such as verb phrases and idioms are particularly challenging for learners to acquire (Nesselhauf & Tschichold, 2002). In fact, both Zhang (1993) and Sugiura (2002) conclude that the 'unnaturalness' of language learners' sentence structures points to a lack of collocational knowledge of English.

It is also interesting to note here that other corpus findings with regard to the types of 'general nouns' used most frequently in spoken registers such as journalistic interviews (Butler, 1998; Mahlberg, 2005) did not show much overlap with the specific nouns found in this focus on the formal, written register, further demonstrating the

essential differences in spoken and written registers. Both Butler (1998) and Mahlberg (2005) found very high-frequency nouns referring to people in their speech-heavy corpora, none of which were found in this study, which focuses on less quotidian, more informational discourse.

5.2 Register Awareness

Much of academic writing teaching, both for native and non-native speakers, consists of raising students' awareness of the formal academic register they should employ in composition writing without denigrating the beauty of the variation inherent to their speech. One general outcome of large-scale corpus studies is that spoken and written language can be described as quantitatively different in their respective uses of particular word classes, even within the same genre such as academic discourse (Biber et al., 1999; Byrd & Reid, 1998). The use of function words associated with complex noun phrases such as articles and prepositions is particularly indicative of formal, academic writing.

> As differences [among text types] are less marked with coordinators and subordinators than with the function words that operate specifically at the phrase level, it seems justified to conclude that register differences are more connected with the build-up of phrases than with the connection of clauses. (Biber et al., 1999, p. 93)

In comparing different genres such as journalistic writings and fiction with that of academic articles and textbooks, corpus studies have also demonstrated that the use of particular language structures differs depending on the genre. Conversation and fiction, as more 'involved' and 'interactional' forms of language, utilize a greater proportion of pronouns, whereas

> prepositions generally seem to be of slightly higher rank in the academic frequency list, reflecting the importance of logical relationships in academic writing…and the prevalence of noun-phrase post-modification using prepositional phrases. (Carter & McCarthy, 2006 as cited in O'Keeffe, McCarthy, & Carter, 2007, p. 201)

This quantitative difference was the driving force behind the present focus of this study on N + P clusters in college-level composition. Native speakers utilize N + P clusters in great numbers, and non-native speakers, in their efforts to emulate the formal, academic register, should also. Because of the 'complex subject matter' of such writing and its 'high informational load,' a higher lexical density, especially with regard to nouns (Biber et al., 1999, p.117), is required of college composition writers.

By focusing on only those N + P clusters with the highest frequencies and exhibiting very robust attractions, we can isolate those structures that are quite restricted by the grammar of English while also being much more common than the relatively fixed idiomatic expressions at one end of Sinclair's grammar continuum. Sinclair (1991b) contended that these are just the types of structures most needed by and difficult for learners, whereas learners tend to focus on more generalizable rules at the open end of the continuum as noted by Pawley and Syder (1983):

> It is a characteristic error of the language learner to assume that an element in the expression may be varied according to a phrase structure or transformational rule of some generality, when in fact the variation (if any) allowed in nativelike usage is much more restricted. The result, very often, is an utterance that is grammatical but unidiomatic, e.g. 'You are pulling my legs.' (p. 215)

A look back at some of the learner errors found in this study brings this point home: **revenge with someone, *opinion in the situation,*life on danger, *interest about something,*decision for the place, *skills on math, *article of newspapers*.

The appropriate preposition selections here are more restricted and opaque in meaning; a simple semantic explanation would fall short.

5.3 Cohesion in Rhetoric: The Role of Prepositions

A lexico-grammatical approach entails that we take advantage of the frequently occurring phrasal units that we can now get access to quite easily through the application of concordancing software programs to massive amounts of running text representing actual language use. In this approach, we can essentially ignore the spaces on the page that occur between words because these spaces have no place in the mind, nor in speech, nor in the communication of ideas. Halliday and Hasan's seminal work on *Cohesion in English* (1976) succeeded in outlining the many structural forms that cohesion in discourse can take. Connor (1984), Scarcella (1984), and Hinkel (2004) have followed up extensively on cohesion in academic writing, especially with regard to learner and native speaker differences. However, the role of prepositions and N + P clusters in phrase-level cohesion has been largely overlooked.

In his introductory linguistics textbook, Gee (1993) includes a final chapter on discourse as language in context, in which he provides an excellent example of the many ways that cohesion (and thereby greater coherence) can be achieved within a span of just two sentences. According to Gee, the six major classes of cohesive devices are anaphoric pronouns, determiners and quantifiers, conjunctions, substitution, ellipsis, and lexical cohesion (p. 410). We should add to this list the category of prepositions, which always serve to link their object noun phrases to other words in a sentence. Furthermore, the choice of which preposition to use depends essentially on the choice of words surrounding it.

Given the significant contribution that prepositions have been shown to make to the juncture of nouns and their adjectival post-modifiers in written academic discourse (rhetoric) and the importance of developing sophisticated academic writing skills for students' higher education pursuits, it behooves us to pay more attention to helping our students develop better writing (and reading) habits at the phrasal level in their assimilation and construction of coherent English sentences.

In essence, prepositions serve to hold sentences together at the phrase level, much like coordinating and subordinating conjunctions hold them together at the clause level, and phrasal sentence connectors, in which prepositions again play a major role, serve many functions in holding sentences and paragraphs together at the discourse level. For this reason, it is difficult to understand why prepositions have been left out of extended discussions on the various ways to achieve cohesion in academic writing (for example, in Gee, 1993; Halliday & Hasan, 1976; Schiffrin, 2006). Reid (1988) does, however, include prepositions in her category of coherence variables, and although cohesion and coherence are not the same thing, cohesion does tend to add to the coherence of a piece of writing. In fact, cohesion is one of the main criteria for the evaluation of college-level essays, and prepositions certainly play a role here (Biber, 1986). When a non-native writer uses an inappropriate preposition or fails to use one where required by the standard grammar of the language, the sentence is stilted, which may obscure meaning or simply draw unnecessary attention on the part of the reader(s) to the anomaly. Reid (1988) contends that "prepositional phrases in written discourse are an indicator of syntactic maturity and complexity" (p. 81). Non-native speakers would benefit from this type of textual knowledge.

The role of N + P clusters in the general cohesion of academic writing has not been directly targeted nor fully explored. Schmid (2000) examines 'shell nouns,' which he describes as abstract nouns followed by a *that*-clause, a *wh*- clause, or a *to* infinitive such as in *The fact that I have no job*. Hunston and Francis (2000) discusses the role of 'shell nouns' in corpora of academic writing. Also, the function of such nouns in cohesion in written texts by both non-native speakers and published writers is examined in Aktas and Cortes (2008). This and other disparate research such as Francis (1986) on 'anaphoric nouns' and (1994) on 'labelling nouns,' Ivanic (1991) on 'carrier nouns,' Flowerdew (2003, 2006) on 'signalling nouns' and Mahlberg (2005) on frequent 'general nouns' having "local textual functions" (p. 3) need to be reviewed and consolidated in light of N + P clusters as those common nouns that appear to be functioning as lexical units framing other nouns and that may also contribute to textual cohesion at the phrase level in academic discourse.

5.4 Pedagogical Implications: Corpus-Informed Language Teaching (CILT)

In light of frequency-based approaches to language description, much research has been done in the area of corpus-informed language teaching and data-driven learning (Johns, 1994; Nesselhauf, 2004a; Partington, 1998; Scott & Tribble, 2006; Sinclair, 1991b, 1999, 2004; Tribble, 2001). An underlying assumption of applying corpus-based findings to language teaching is that frequent language structures for native speakers equal useful structures for language learners.

As Aarts (1991) points out, traditional grammars have been intuition-based, and recent technological capabilities have allowed for the rapid development of more observation-based grammars. In other words, the rational/empirical pendulum in applied

linguistics can now swing back towards a greater focus on actual language behavior rather than on native speaker competence as a primary source of information for language pedagogy. Language is inherently social, and meaning is defined by usage. "If meaning is defined as use, frequency is part of the meaning of words" (Mahlberg, 2005, p. 36). The fact that particular forms are used frequently, which can be established through empirical corpus inquiry, indicates the general range of meanings for those forms and their general utility in certain registers.

Teachers can now consult a massive amount of research based on corpus analysis in order to validate (or not) their deeply held assumptions about the way the English language works. Those assumptions and intuitions are based on specific experience, and we tend to notice the unusual more than the common, whereas now we can base our ideas on massive accumulations of actual native-speaker and learner language use. The relevance of language corpora findings to the teaching of language as used by native speakers cannot be overstated (McEnery & Wilson, 1997; Hung, T. T. N., 2002).

> In the past few decades, there has been an unhealthy dichotomization of form-focused instruction and meaning-focused instruction. Corpus studies have shown that linguistic forms, contexts, and meanings are inextricably linked…the co-occurrence of lexical items in different contexts is crucial to the meanings that they take on and the pragmatic functions that they perform. The engagement of teachers in corpus enquiry will help them to gain a better understanding of the relationship between form and meaning, which can in turn redress the balance between form and meaning in the language curriculum. (Tsui, 2005, p. 352)

Several academic ESL vocabulary textbooks that have been designed from corpus frequencies are those by Bunting (2006), Dingle (2008), Jones (2004), Schmitt and Schmitt (2005) and Woolard (2004). All of these works are based on frequently-occurring lexical items such as those provided in Coxhead's Academic Word List (2000). A lexico-grammatical approach recognizes that these content words occur frequently in

phrasal patterns in academic writing, patterns that can be discerned from careful corpus study as demonstrated above. In fact, such research is now being undertaken by Coxhead and Byrd (forthcoming) on the most frequent two-word clusters involving the content words from the Academic Word List. These researchers are already finding many strong relationships between nouns and their post-modifying prepositions (Byrd, personal communication) as presented in this study. These kinds of empirically-based teaching resources are sure to become more widely available to us as the technology becomes more widespread, and students and teachers could benefit from using corpus-based textbooks from ESL/EFL publishers and materials writers. In fact, Howarth (1998) notes that

> a glance through recent [at the time of writing] EFL coursebooks…shows that teachers and materials writers are paying increasing attention to the necessity of learners to acquire knowledge of collocations and are aware that this component of competence should be addressed explicitly. Although this need was recognized and examined in detail as long ago as the 1930s…, the prolonged influence of generative grammar and the purer forms of communicative language teaching downgraded vocabulary learning in the syllabus and made teachers and applied linguists shy away from any materials that smacked of phrasebook learning. (p. 30)

As for the presentation of N + P cluster frequencies to ESL/EFL students, they should first and foremost be given lists of such lexical units and be encouraged to 'notice' them in contexts in their academic reading (Lewis, 2000). The utility of phrase lists to language study was largely abandoned (along with audio-lingual methods) with the advent of more communicative language teaching methods. However, students in language learning classes very frequently make their own lists as a method of making the study of vocabulary and its retention more efficient. However, Coxhead (2000) cautions against simply relying on word lists for teaching academic vocabulary:

> The AWL [Academic Word List] is the result of a corpus-based study. Such studies create lists, concordances, or data concerning the clustering of linguistics items in coherent, purposeful texts. The use of this research method, however, does not imply that language teaching and learning should rely on decontextualised methods. Instead, the AWL might be used to set vocabulary goals for EAP courses, construct relevant teaching materials, and help students focus on useful vocabulary items. (p. 227)

Clearly, word/phrase lists also need to be contextualized for learners in order to become more pedagogically useful.

Both Sinclair's and Biber's corpus work has resulted in the production of comprehensive reference grammars for students of English, the Collin's Cobuild series (1991a) and the Longman English Grammar (1999) respectively. But there is more work to be done, and with our current technological capacity to process huge amounts of information in a matter of seconds, work that used to take years in the creation of comprehensive dictionaries, now makes it possible for us to teach English grammar and lexis in unison as native speakers actually use it in various registers. In discussing the lexical syllabus for language learning, Sinclair and Renouf (1988) recommend that "for any learner of English, the main focus of study should be on (a) the commonest word forms in the language; (b) their central patterns of usage [and]; (c) the combinations which they typically form" (p. 148).

Some would claim that 'local' errors such as with prepositions are not worthy of much attention in the second language writing classroom because they have little effect on the transfer of meaning. However, second language learners want to be corrected on every point so that their writing is accurate and not stigmatized by distinctively non-native usage. Errors with regard to the small, function words such as articles, prepositions, and conjunctions are quite noticeable to native speakers and also identifying features of non-native prose and speech. Language learners want to be accurate in their

English language usage, which can be better accomplished with some focus on form and on recognizable patterns.

Certain vocabulary items specific to particular disciplines would become more frequent and therefore more relevant to teaching students in particular disciplines in the content areas. This bodes well for applications to the learning of topical vocabulary in English for Specific Purposes (ESP). Depending on the content area, such as law, medicine, business, history, or science, topic-specific vocabulary frequencies would become more prevalent in relevant texts. Indeed, even some freshman composition courses today are focused on particular themes based on students who have declared a major. Corpus linguistics is a promising area of research for the enhancement of higher education experiences that are also relevant to students' specific discipline choices.

Second language teachers who have little time for research should seek out materials that use the discoveries and implications of empirical corpus studies to inform their curriculum and ELT materials design. Language is constantly in a state of flux, and we now have at our fingertips a way to capture a piece of the picture distinctly focused on particular text types and particular topics. The potential for EAP/ESP courses to be designed around vocabulary frequencies, as can be discovered through the use of a concordancer, opens new opportunities for students to prepare themselves for their future work.

With regard to specific applications in the classroom, Coxhead (2008) employs "three psychological conditions of noticing, retrieval, and generation" (p. 156). The first step, 'noticing,' is achieved by making students aware of formulaic sequences in academic reading activities by highlighting them. 'Retrieval' refers to the need for repeated exposure to formulaic sequences through the "retelling of key sections of source

texts," (p.156), the utilization of 'word cards,' and classroom 'recycling.' The researcher's writing students have made their own laminated bookmarks out of frequent phrase lists, especially N + P clusters, culled from a content area textbook. 'Generation' involves "isolating target collocations in sentences and creating new texts around them" (p. 156). Target items in source texts can be manipulated by "paraphrasing, summary writing, and quotation practice" (p. 156).

5.5 Implications for Future Research

Of course, there are many more two-word N + P clusters with absolute frequencies below twenty in the Corpus under analysis here (see Appendix L) such as *factor in* for 33% (15/45) and *advances in* for 48% (14/29). Although they are below the frequency cutoff rate for this study of 20 tokens per 500,000 words, the ratio of each noun's occurrence with a particular preposition may be quite high when considered relative to input probabilities for the preposition in question. This prospect implicates the need for further investigations of N + P cluster frequencies, and the results above represent only a preliminary consideration. A diachronic study of learner usage of N + P clusters is also warranted with an eye toward effective teaching methodologies.

The AntConc freeware concordancer software program used in this research was specifically designed with a user-friendly interface by its creator for use in the L2 classroom (Anthony, 2004). Data-driven learning as advocated by Johns (1994, see also Scott & Tribble, 2006; Thurston, 1997; Thurston & Candlin, 1998; Tribble & Jones, 1998) offers a way to address both grammar and vocabulary simultaneously using concordancer technology in the classroom in order to easily discover frequent collocations in use. L2 teachers should begin to maintain their own archives of student

writing in the form of independent monitor corpora,[30] which can serve as an excellent resource for error analysis, revision and editing practices, and diachronic development.

We should think of vocabulary as individual lexical items no more than we think of words as their individual letters or sounds. The growth in the number of grammatical tagging categories (as in the CLAWS8 POS Tagger) demonstrates the finer distinctions that need to be made in the actual present-day usage of words and obliterates the traditional hard lines drawn between and among word classes. Grammatical categories are no more static than vocabulary. The preposition *of* in particular is used in a variety of ways other than in the genitive construction, and its particular range of use calls into question its relegation to this confining a category (Sinclair, 1991b). N + P clusters as demonstrated above are viable and useful units in the construction of English sentences. The inherent inseparability of grammar and vocabulary is a promising area for corpus-based studies in a lexico-grammatical approach to actual language usage. With the modern availability of corpus data, we no longer need to rely on outdated grammars nor on our own personal and frequently faulty impressions of how the English language works. Language patterns represent the interface between grammar and lexis, and, here, frequency matters. "If we examine the frequency of words in a large corpus of English, a picture emerges where the first 2,000 or so word-forms do most of the work, accounting for more than 80% of all of the words in spoken or written text" (O'Keeffe, McCarthy, & Carter, 2007, p. 32). So language learners are well-served by giving them lots of exposure to what they really need: a hard-working, core vocabulary with some relevant

---

[30] Of course, this kind of cataloguing is subject to IRB guidelines with regard to research involving human subjects and should only be done with participant anonymity, understanding, and agreement.

discussion of their embedded forms, distribution of use, respective functions in discourse, and topical contextualization.

References

Aijmer, K., & Altenberg, B. (1991). *English corpus linguistics: Studies in honour of Jan Svartvik*. London; New York: Longman.

Aktas, R. N., & Cortes, V. (2008). Shell nouns as cohesive devices in published and ESL student writing. *Journal of English for Academic Purposes, 7*(1), 3-14.

Allerton, D. J., Nesselhauf, N., & Skandera, P. (2004). *Phraseological units: Basic concepts and their application*. Basel: Schwabe.

Anthony, L. (2008). *AntConc 3.2.2w for Windows*. Waseda, Japan: Retrieved from http://www.antlab.sci.waseda.ac.jp/

Anthony, L. (2008). *Laurence Anthony's Homepage.* Retrieved July 1, 2008, from http://www.antlab.sci.waseda.ac.jp/

Anthony, L. (2008). *Laurence Anthony's Homepage - Software.* Retrieved July 1, 2008, from http://www.antlab.sci.waseda.ac.jp/software.html

Anthony, L. (2004). AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning, December 10th, 2004, Waseda University, Tokyo,*

Azar, B. S. (2003). *Fundamentals of English grammar* (3rd ed.). White Plains, NY: Pearson Education.

Azevedo, M. M. (1980). The interlanguage of advanced learners: An error analysis of graduate students' Spanish. *IRAL, International Review of Applied Linguistics in Language Teaching, 18*(3), 217-227.

Barber, C. L. (1993). *The English language: A historical introduction*. Cambridge; New York: Cambridge University Press.

Benson, B., Deming, M. P., Denzer, D., & Valeri-Gold, M. (1992). A combined Basic Writing/English as a second language class: Melting pot or mishmash? *Journal of Basic Writing, 11*(1), 58-74.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam; Philadelphia: John Benjamins Publishing Co.

Biber, D., & Clark, V. (2000). Historical shifts in modification patterns with complex noun phrase structures: How long can you go without a verb? *English Historical*

Syntax and Morphology: Selected Papers from 11 ICEHL, Santiago de Compostela, 7-11 September 2000, 43-66.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselggård, & S. Oksefjell (Eds.), *Out of Corpora: Studies in honour of Stig Johansson* (pp. 181-190). Amsterdam: Rodopi.

Biber, D., Conrad, S., & Cortes, V. (2003). Lexical bundles in speech and writing: An initial taxonomy. In G. N. Leech, T. McEnery, P. Rayson & A. Wilson (Eds.), *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech* (pp. 71-92). New York: Peter Lang.

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., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly, 36*(1), 9-48.

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

Biber, D. (1985). Investigating macroscopic textual variation through multifeature/ multidimensional analyses. *Linguistics, 23*(2), 337-360.

Biber, D. (1986). On the investigation of spoken/written differences. *Studia Linguistica, 40*(1), 1-21.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263-286.

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing, 14*(3), 191-205.

Boers, F., & Demecheleer, M. (1998). A cognitive semantic approach to teaching prepositions. *ELT Journal, 52*(3), 197-204.

Brinton, L. J., & Traugott, E. C. (2005). *Lexicalization and language change*. New York: Cambridge U Press.

Brown Corpus. (2008). Retrieved February 14, 2008, from http://en.wikipedia.org/wiki/Brown_Corpus

Bunting, J. B. (2006). *College vocabulary: English for academic success*. Boston, MA: Thomson Heinle.

Burger, H., Dobrovol'skij, D., Kuhn, P., & Norrick, N. R. (Eds.). (2007). *Phraseology: An international handbook of contemporary research*. Berlin: Walter de Gruyter.

Butler, C. S. (1998). Collocational frameworks in Spanish. *International Journal of Corpus Linguistics, 3*(1), 1-32.

Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, J. (2002). Phonological evidence for exemplar storage of multiword sequences. *Studies in Second Language Acquisition, 24*(2), 215-221.

Byrd, P. (1998). Rethinking grammar at various proficiency levels: Implications of authentic materials for the EAP curriculum. In J. M. Reid, & P. Byrd (Eds.), *Grammar in the composition classroom* (pp. 69-97). New York: Heinle & Heinle Publishers.

Campion, M., & Elley, W. (1971). *An academic vocabulary list*. Wellington, NZ: New Zealand Council for Educational Research.

Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English: A comprehensive guide*. Cambridge, UK; New York: Cambridge University Press.

Coe, N. (2001). Speakers of Spanish and Catalan. In M. Swan, & B. Smith (Eds.), *Learner English: A teacher's guide to interference and other problems* (2nd ed., pp. 90-112). Cambridge: Cambridge University Press.

Coffin, S., & Hall, B. (1998). *Writing workshop: A manual for college ESL writers*. New York, NY: McGraw-Hill Companies, Inc.

Connor, U. (1984). A study of cohesion and coherence in English as a second language student's writing. *Papers in Linguistics, 17*, 301-316.

Connor, U., Nagelhout, E., & Rozycki, W. V. (Eds.). (2008). *Contrastive rhetoric: Reaching to intercultural rhetoric*. Amsterdam: John Benjamins Publishing Co.

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

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*(4), 397-423.

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

Coulthard, M. (Ed.). (1994). *Advances in written text analysis*. London: Routledge.

Cowan, R. (2008). *The teacher's grammar of English: A course book and reference guide*. New York; Cambridge: Cambridge University Press.

Coxhead, A. (1998). *An academic word list*. Wellington, NZ: Victoria University of Wellington.

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

Coxhead, A. (2002). The academic word list: A corpus-based wordlist for academic purposes. *Teaching and Learning by Doing Corpus Analysis: Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz 19-24 July, 2000, , 42* 73-85.

Coxhead, A. (2008). Phraseology and English for academic purposes: Challenges and opportunities. In S. Granger, & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 149-161). Amsterdam; Philadelphia: John Benjamins Publishing Co.

Coxhead, A., & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing, 16*(3), 129-147.

Coxhead, A., & Byrd, P. (forthcoming). *The AWL: Collocations and recurrent phrases*. Ann Arbor, MI: Michigan University Press.

Desmet, C., & Balthazor, R. (2005). *Finding patterns in textual corpora: Data mining, research, and assessment in first-year composition.* Unpublished manuscript. Retrieved August 13, 2008, from http://www.english.uga.edu/emma/DesmetandBalthazorCW05.pdf

Dingle, K. (2008). *Vocabulary power 3: Practicing essential words*. White Plains, NY: Pearson Education, Inc.

Dykstra, P. (1997). The patterns of language: Perspectives on teaching writing. *Teaching English in the Two-Year College, 24*(2), 136-144.

Ene, S. E. (2006). The last stages of second language acquisition: Linguistic evidence from academic writing by advanced non-native English speakers. Unpublished doctoral dissertation, University of Arizona, Tucson, Arizona.

Erman, B. (2007). Cognitive processes as evidence of the idiom principle. *International Journal of Corpus Linguistics, 12*(1), 25-53.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text, 20*(1), 29-62.

Feigenbaum, S., & Kurzon, D. (2002). *Prepositions in their syntactic, semantic, and pragmatic context*. Amsterdam: John Benjamins Publishing Co.

Ferris, D. (2002). *Treatment of error in second language student writing*. Ann Arbor: University of Michigan Press.

Firth, J. R. (1957). *Papers in linguistics, 1934-1951*. London; New York: Oxford University Press.

Flower, J. (2002). *Phrasal verb organizer with mini-dictionary*. Boston, MA: Thomson.

Flowerdew, J. (2003). Signalling nouns in discourse. *English for Specific Purposes, 22*(4), 329-346.

Flowerdew, J. (2006). Use of signalling nouns in a learner corpus. *International Journal of Corpus Linguistics, 11*(3), 345-362.

Flowerdew, J., & Li, Y. (2007). Plagiarism and second language writing in an electronic age. *Annual Review of Applied Linguistics, 27*, 161-183.

Francis, G. (1986). *Anaphoric nouns*. Birmingham: English Language Research.

Francis, G. (1994). Labelling discourse: An aspect of nominal-group lexical cohesion. In M. Coulthard (Ed.), *Advances in written text analysis* (pp. 84-101). London: Routledge.

Francis, G., Hunston, S., & Manning, E. (Eds.). (1998). *Grammar patterns 2: Nouns and adjectives*. London: Harper Collins Publishers.

Francis, W. N., Kučera, H., & Mackie, A. W. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.

Fries, C. C. (1952). *The structure of English: An introduction to the construction of English sentences*. New York: Harcourt Brace.

Fries, C. C. (1954). *Teaching and learning English as a foreign language*. Ann Arbor: University of Michigan Press.

Fries, C. C., & Traver, A. A. (1965). *English word lists*. Washington, D.C.: American Council on Education.

Gee, J. P. (1993). *An introduction to human language: Fundamental concepts in linguistics*. Upper Saddle River, NJ: Prentice-Hall Inc.

Germany, P., & Cartes, N. (1995). Spatial prepositions in English as a foreign language: One aspect of interlanguage. [Preposiciones espaciales del ingles como lengua

extranjera: Un aspecto de interlengua] *Estudios De Linguistica Aplicada, 13*(21-22), 44-55.

Ghadessy, M., Henry, A., Roseberry, R. L., & Sinclair, J. M. (Eds.). (2001). *Small corpus studies and ELT: Theory and practice*. Amsterdam: John Benjamins Publishing Co.

Gitsaki, C. (1999). *Second language lexical acquisition: A study of the development of collocational knowledge*. San Francisco: International Scholars Publications.

Granger, S., Hung, J., & Petch-Tyson, S. (2002). *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins Publishing Co.

Granger, S., & Meunier, F. (2008). *Phraseology: An interdisciplinary perspective*. Amsterdam; Philadelphia: John Benjamins Publishing Co.

Halliday, M. A. K. (1989). Some grammatical problems in scientific English. *Australian Review of Applied Linguistics, 6*, 13-37.

Halliday, M. A. K. (1991). Corpus studies and probabilistic grammar. In K. Aijmer, & B. Altenberg (Eds.), *English corpus linguistics: Studies in honour of Jan Svartvik* (pp. 30-43). London: Longman.

Halliday, M. A. K. (2004a). *Introduction to functional grammar* (3rd revised by Matthiessen, C.M.I.M. ed.). London: Arnold.

Halliday, M. A. K. (2004b). *Lexicology and corpus linguistics: An introduction*. London: Continuum.

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

Harris, M., & Silva, T. (1993). Tutoring ESL students: Issues and options. *College Composition and Communication, 44*(4), 525-537.

Hasselggård, H., & Oksefjell, S. (Eds.). (1999). *Out of corpora: Studies in honour of Stig Johansson*. Amsterdam: Rodopi.

Hasselgren, A. (2002). Learner corpora and language testing: Small words as markers of learner fluency. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 143-173). Amsterdam, Netherlands: John Benjamins Publishing Co.

Hemchua, S., & Schmitt, N. (2006). An analysis of lexical errors in the English compositions of Thai learners. *Prospect, 21*(3), 3-25.

Henning, G. H. (1978). A developmental analysis of errors of adult Iranian students of English as a foreign language. *Language Learning, 28*(2), 387-397.

Hinkel, E. (2004). *Teaching academic ESL writing: Practical techniques in vocabulary and grammar.* Mahwah, NJ: Lawrence Erlbaum Associates.

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

Howarth, P. A. (1998). Phraseology and second language proficiency. *Applied Linguistics, 19*(1), 24-44.

Hudson, R. (1994). About 37% of word-tokens are nouns. *Language, 70*(2), 331-339.

Hung, T. T. N. (2002). The use of language corpora in the teaching of English. *Hong Kong Journal of Applied Linguistics, 7*(1), 34-48.

Hunston, S., & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins Publishing Co.

Ivanic, R. (1991). Nouns in search of a context: A study of nouns with both open- and closed-system characteristics. *IRAL, 29*(2), 93-114.

Jiménez-Catalán, R. M. (1996). Frequency and variability in errors in the use of English prepositions. *Miscelanea, 17*, 171-187.

Johns, T. (1994). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Odlin (Ed.), *Perspectives on Pedagogical Grammar* (pp. 293-313). Cambridge: Cambridge University Press.

Jones, B. (2004). *Academic word power*. Boston, MA: Thomson Heinle.

Kennedy, G. D. (1991). *Between* and *through*: The company they keep and the functions they serve. In K. Aijmer, & B. Altenberg (Eds.), *English corpus linguistics: Studies in honour of Jan Svartvik* (pp. 95-127). London: Longman.

Kennedy, G. D. (2003). *Structure and meaning in English: A guide for teachers*. Harlow, UK: Pearson Longman.

Khampang, P. (1974). Thai difficulties in using English prepositions. *Language Learning, 24*(2), 215-222.

Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL Quarterly, 26*(1), 81-112.

Kolln, M., & Funk, R. (2006). *Understanding English grammar* (7th ed.). New York: Pearson Education, Inc.

Koosha, M., & Jafarpour, A. A. (2006). Data-driven learning and teaching collocation of prepositions: The case of Iranian EFL adult learners. *Asian EFL Journal, 8*(8), 192-209.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.

Lafford, B. A., & Ryan, J. M. (1995). The acquisition of lexical meaning in a study abroad context: The Spanish prepositions *por* and *para*.. *Hispania, 78*(3), 528-547.

Lancaster University Centre for Computer Corpus Research on Language. (2008). *CLAWS part-of-speech tagger for English.* Retrieved June 11, 2008, from http://ucrel.lancs.ac.uk/claws/

Lee, I. (2004). Error correction in L2 secondary writing classrooms: The case of Hong Kong. *Journal of Second Language Writing, 13*(4), 285-312.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English (based on the British National Corpus)*. Harlow, England: Longman.

Levy, S. A. (2004). Lexical bundles in professional and student writing. Unpublished doctoral dissertation, University of the Pacific, Stockton, CA.

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

Lewis, M. (1997). *Implementing the lexical approach: Putting theory into practice.* Hove, England: Language Teaching Publications.

Lewis, M. (2000). *Teaching collocation: Further developments in the lexical approach*. Hove, England: Language Teaching Publications.

Li, X. (2008). From contrastive rhetoric to intercultural rhetoric: A search for collective identity. In U. Connor, E. Nagelhout & W. V. Rozycki (Eds.), *Contrastive rhetoric: Reaching to intercultural rhetoric.* (pp. 11-24). Amsterdam: John Benjamins Publishing Co.

Lindstromberg, S. (1998). *English prepositions explained*. Amsterdam: John Benjamins Publishing Co.

Mahlberg, M. (2005). *English general nouns: A corpus theoretical approach*. Amsterdam: John Benjamins Publishing Co.

McCarthy, M., & O'Dell, F. (2004). *English phrasal verbs in use: Intermediate*. Cambridge: Cambridge University Press.

McCarthy, M., & O'Dell, F. (2007). *English phrasal verbs in use: Advanced*. Cambridge: Cambridge University Press.

McEnery, T., & Wilson, A. (1997). Teaching and language corpora (TALC). *ReCALL, 9*(1), 5-14.

Meziani, A. (1984). Moroccan learners' English errors: A pilot study. *IRAL, International Review of Applied Linguistics in Language Teaching, 22*(4), 297-309.

Miller, T. (Ed.). (2007). *How I learned English*. Washington, DC: National Geographic Society.

Moon, R. (2007). Corpus linguistic applications with English corpora. In H. Burger, D. Dobrovol'skij, P. Kuhn & N. R. Norrick (Eds.), *Phraseology: An international handbook of contemporary research* (pp. 1045). Berlin: Walter de Gruyter.

Moreira-Rodriguez, A. (2006). 'The book on the table,' 'the man in the moon': Post-modification of nouns by preposition + noun in English and Castilian. *Bulletin of Spanish Studies, 83*(1), 53-72.

Morenberg, M. (2002). *Using grammar*. New York: Oxford University Press Inc.

Moreno, A. I. (2008). The importance of comparable corpora in cross-cultural studies. In U. Connor, E. Nagelhout & W. V. Rozycki (Eds.), *Contrastive rhetoric: Reaching to intercultural rhetoric* (pp. 25-41). Amsterdam: John Benjamins Publishing Co.

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

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

Neff, J., Ballesteros, F., Dafouz, E., Martinez, F., & Rica, J. (2004). A contrastive functional analysis of errors in Spanish EFL university writers' argumentative text: Corpus-based study. In E. Fitzpatrick (Ed.), *Corpus linguistics beyond the word: Corpus research from phrase to discourse.* (pp. 203-225). Amsterdam: Rodopi.

Nesselhauf, N. (2004a). Learner corpora and their potential for language teaching. In J. M. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp. 125-152). Amsterdam: John Benjamins Publishing Co.

Nesselhauf, N. (2004b). What are collocations? In D. J. Allerton, N. Nesselhauf & P. Skandera (Eds.), *Phraseological units: Basic concepts and their application* (pp. 1-21). Basel: Schwabe.

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

O'Dowd, E. M. (1998). *Prepositions and particles in English: A discourse-functional account*. New York: Oxford University Press.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.

Oshima, A., & Hogue, A. (2005). *Writing academic English* (6th ed.) Pearson Longman.

Partington, A. (1998). *Patterns and meanings: Using corpora for English language research and teaching*. Amsterdam: John Benjamins Publishing Co.

Pawley, A., & Syder, F. H. (1983). Natural selection in syntax: Notes on adaptive variation and change in vernacular and literary grammar. *Journal of Pragmatics, 7*(5), 551-579.

Praninskas, J. (1972). *American university word list*. London: Longman.

Pu, J. (2003). Colligation, collocation, and chunk in ESL vocabulary teaching and learning. *Foreign Language Teaching and Research, 35*(6), 438-445.

Raimes, A. (2004). *Grammar troublespots: A guide for student writers* (3rd ed.). Cambridge: Cambridge University Press.

Reid, J. M. (1988). Quantitative differences in English prose written by Arabic, Chinese, Spanish, and English students. Unpublished doctoral dissertation, Colorado State University, Fort Collins.

Reid, J. M., & Byrd, P. (1998). *Grammar in the composition classroom*. New York: Heinle & Heinle Publishers.

Renouf, A., & Sinclair, J. M. (1991). Collocational frameworks in English. In K. Aijmer, & B. Altenberg (Eds.), *English corpus linguistics: Studies in honour of Jan Svartvik* (pp. 128-144). London: Longman.

Reppen, R., Fitzmaurice, S. M., & Biber, D. (2002). *Using corpora to explore linguistic variation*. Amsterdam; Philadelphia: John Benjamins Publishing Co.

Richards, J. C., Platt, J. T., & Platt, H. K. (1992). *Longman dictionary of language teaching and applied linguistics* (2nd ed.). Essex, England: Longman.

Salem, I. (2007). The lexico-grammatical continuum viewed through student error. *ELT Journal, 61*(3), 211-219.

Scarcella, R. C. (1984). Cohesion in the writing development of native and non-native English speakers. Unpublished doctoral dissertation, University of Southern California, Los Angeles.

Schiffrin, D. (2006). Discourse. In R. W. Fasold, & J. Connor-Linton (Eds.), *An Introduction to Language and Linguistics* (pp. 169-203). Cambridge: Cambridge University Press.

Schmid, H. (2000). *English abstract nouns as conceptual shells: From corpus to cognition*. Berlin; New York: Mouton de Gruyter.

Schmitt, D., & Schmitt, N. (2005). *Focus on vocabulary: Mastering the academic word list.* White Plains, NY: Pearson Education Inc.

Schmitt, N. (1997). *Vocabulary: Description, acquisition, pedagogy*. Cambridge: Cambridge Language Teaching Library.

Schmitt, N. (2000). *Vocabulary in language teaching.* Cambridge: Cambridge University Press.

Schmitt, N. (2004). *Formulaic sequences: Acquisition, processing, and use*. Amsterdam: John Benjamins Publishing Co.

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins Publishing Co.

Sinclair, J. M. (Ed.). (1991a). *Collins COBUILD English guides 1: Prepositions*. London: Harper Collins Publishers.

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

Sinclair, J. M. (1999). A way with common words. In H. Hasselggård, & S. Oksefjell (Eds.), *Out of Corpora: Studies in honor of Stig Johansson* (pp. 157-179). Amsterdam: Rodopi.

Sinclair, J. M. (2004). *How to use corpora in language teaching*. Amsterdam; Philadelphia: John Benjamins Publishing Co.

Sinclair, J. M., & Carter, R. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.

Sinclair, J. M., & Renouf, A. (1988). A lexical syllabus for language learning. In R. Carter, & M. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 140-160). London: Longman.

Sosa, A. V., & MacFarlane, J. (2002). Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word *of.*. *Brain and Language, 83*(2), 227-236.

Stevens, V. (1991). Classroom concordancing: Vocabulary materials derived from relevant, authentic text. *English for Specific Purposes, 10*(1), 35-46.

Stockwell, R. P., Bowen, J. D., & Martin, J. W. (1965). *The grammatical structures of English and Spanish*. Chicago: University of Chicago Press.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Malden, Mass.: Blackwell Publishers.

Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics, 7*(2), 215-244.

Stubbs, M. (2004). A quantitative approach to collocations. In D. J. Allerton, N. Nesselhauf & P. Skandera (Eds.), *Phraseological units: Basic concepts and their application* (pp. 107-119). Basel: Schwabe.

Sugiura, M. (2002). Collocational knowledge of L2 learners of English: A case study of Japanese learners. In T. Saito, J. Nakamura & S. Yamazaki (Eds.), *English Corpus Linguistics in Japan* (pp. 303-323). Amsterdam: Rodopi.

Swan, M., & Smith, B. (Eds.). (2001). *Learner English: A teacher's guide to interference and other problems* (2nd ed.). Cambridge: Cambridge University Press.

Swick, E. (2005). *Practice made perfect: English pronouns and prepositions*. New York, NY: McGraw Hill Companies.

The University of Georgia. (2007). *Academic Life.* Retrieved March 16, 2008, from http://www.admissions.uga.edu/4_fy_closerlook.html

The University of Georgia. (2008). *Quick Facts about UGA.* Retrieved January 21, 2008, from http://www.uga.edu/profile/facts.html

Thomas, E. C. (2004). Second language acquisition of prepositions: Functional and substantive features. Unpublished doctoral dissertation, University of Essex, Essex, England.

Thornbury, S. (1999). *How to teach grammar*. Essex, England: Pearson Education Limited.

Thorndike, E. L. (1932). *A teacher's word book of 20,000 words*. New York: Teachers College, Columbia University.

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

Thurstun, J. (1997). Using concordances for the contextual teaching of vocabulary. *EA Journal, 15*(2), 29-37.

Thurstun, J., & Candlin, C. N. (1998). Concordancing and the teaching of the vocabulary of academic English. *English for Specific Purposes, 17*(3), 267-280.

Tribble, C. (2001). Small corpora and teaching writing: Towards a corpus-informed pedagogy of writing. In M. Ghadessy, A. Henry, R. L. Roseberry & J. M. Sinclair (Eds.), *Small corpus studies and ELT: Theory and practice* (pp. 381-408). Amsterdam: John Benjamins Publishing Co.

Tribble, C., & Jones, G. (1997). *Concordances in the classroom: A resource guide for teachers*. Houston, TX: Athelstan Publications.

Tsui, A. B. M. (2005). ESL teachers' questions and corpus evidence. *International Journal of Corpus Linguistics, 10*(3), 335-356.

Tyler, A., & Evans, V. (2003). *Semantics of English prepositions: Spatial scenes, embodied meaning, and cognition.* Cambridge, UK: Cambridge University Press.

West, M. (1953). *A general service list of English words*. London: Longman, Green, and Co.

Whitehorn, J. C., & Zipf, G. K. (1943). Schizophrenic language. *Arch. Neurology & Psychiatry, 49*, 831-851.

Wichmann, A. (1997). *Teaching and language corpora*. London: Longman.

Widdowson, H. G. (1989). Knowledge of language and ability for use. *Applied Linguistics 10*(2), 128-137.

Woolard, G. (2004). *Key words for fluency: Learning and practising the most useful words of English*. Boston, MA: Thomson ELT.

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

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication, 3*(2), 215-229.

Yates, J. (1999). *The ins and outs of prepositions: A guidebook for ESL students*. Hauppauge, NY: Barron's Educational Series, Inc.

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.

Zipf, G. K. (1942). Children's speech. *Science, 96*, 344-345.

Zipf, G. K. (1945a). The repetition of words, time-perspective, and semantic balance. *The Journal of General Psychology, 32*, 127-148.

Zipf, G. K. (1945b). The meaning-frequency relationship of words. *The Journal of General Psychology, 33*, 251-256.

APPENDIX A:
The UGALECT Corpus
First 600 Words
(With prepositions highlighted)

1	31271	the	51	1309	about	101	572	person
2	16295	to	52	1261	also	102	569	two
3	14742	of	53	1235	who	103	568	however
4	13622	and	54	1227	she	104	564	men
5	11340	a	55	1143	do	105	559	get
6	9852	in	56	1118	which	106	551	its
7	8243	that	57	1113	them	107	543	may
8	7932	is	58	1073	how	108	542	your
9	4635	for	59	1072	our	109	533	made
10	4330	are	60	1060	had	110	530	any
11	4237	it	61	1052	so	111	505	american
12	3680	as	62	1019	women	112	499	did
13	3620	not	63	996	only	113	499	him
14	3613	be	64	991	other	114	497	society
15	3587	this	65	977	like	115	494	while
16	3494	they	66	964	time	116	493	then
17	3463	with	67	957	out	117	491	after
18	3456	on	68	919	been	118	490	where
19	3196	was	69	911	world	119	486	does
20	3152	have	70	905	my	120	482	students
21	3043	s	71	900	most	121	476	see
22	2903	their	72	897	could	122	469	now
23	2550	he	73	877	being	123	469	today
24	2344	people	74	868	than	124	464	america
25	2341	i	75	868	up	125	463	well
26	2205	or	76	849	life	126	460	another
27	2043	by	77	835	no	127	452	debate
28	2014	his	78	809	states	128	448	help
29	1919	but	79	800	way	129	443	still
30	1919	has	80	788	should	130	442	become
31	1880	from	81	783	just	131	440	own
32	1871	one	82	772	even	132	439	every
33	1833	an	83	755	war	133	438	same
34	1783	more	84	747	some	134	438	things
35	1776	can	85	743	into	135	436	years
36	1568	would	86	738	very	136	435	both
37	1528	her	87	717	new	137	435	think
38	1503	all	88	704	such	138	427	man
39	1487	you	89	654	make	139	427	want
40	1460	many	90	648	t	140	422	each
41	1458	what	91	633	iraq	141	415	better
42	1456	at	92	632	children	142	413	country
43	1448	will	93	615	different	143	409	know
44	1430	were	94	615	united	144	398	between
45	1420	these	95	611	use	145	397	change
46	1415	we	96	598	over	146	392	child
47	1399	because	97	595	first	147	391	take
48	1373	there	98	594	through	148	390	believe
49	1346	when	99	589	school	149	389	used
50	1322	if	100	585	much	150	388	able

151	388	those	201	278	fact	251	227	food
152	381	feel	202	277	always	252	226	long
153	380	me	203	277	put	253	225	education
154	379	work	204	274	using	254	225	makes
155	378	us	205	273	best	255	224	image
156	377	information	206	268	death	256	223	instead
157	376	computer	207	268	example	257	223	point
158	375	audience	208	268	part	258	223	reason
159	375	need	209	266	against	259	223	seems
160	365	good	210	266	back	260	222	mac
161	364	around	211	265	body	261	222	president
162	363	why	212	263	live	262	221	often
163	361	government	213	263	say	263	220	everyone
164	352	never	214	263	woman	264	220	music
165	352	young	215	262	state	265	219	facebook
166	351	go	216	259	idea	266	218	message
167	348	lives	217	258	making	267	218	three
168	347	day	218	258	video	268	217	everything
169	346	before	219	257	someone	269	217	once
170	346	right	220	252	although	270	215	class
171	346	without	221	250	having	271	215	don
172	345	debates	222	248	known	272	215	whether
173	344	important	223	248	social	273	213	certain
174	336	must	224	246	friends	274	212	sex
175	334	going	225	244	really	275	210	nation
176	330	great	226	243	americans	276	208	given
177	327	order	227	243	issue	277	207	keep
178	325	since	228	243	off	278	207	past
179	323	television	229	242	shows	279	207	thought
180	321	during	230	242	too	280	204	form
181	320	though	231	240	down	281	204	lot
182	318	money	232	239	others	282	201	am
183	315	look	233	237	college	283	201	ways
184	313	public	234	237	health	284	200	old
185	309	parents	235	237	media	285	200	words
186	306	come	236	237	thing	286	198	house
187	305	technology	237	237	trying	287	198	start
188	303	something	238	236	end	288	197	girls
189	301	high	239	236	give	289	196	ad
190	301	themselves	240	235	found	290	196	care
191	294	internet	241	234	away	291	196	less
192	294	place	242	234	times	292	195	says
193	291	year	243	233	problems	293	195	true
194	285	obama	244	231	home	294	194	age
195	284	family	245	230	according	295	194	enough
196	283	problem	246	230	attention	296	193	citizens
197	283	show	247	230	system	297	193	due
198	282	candidates	248	229	human	298	192	doing
199	282	find	249	228	candidate	299	192	few
200	279	said	250	228	love	300	192	history

301	192	play	351	167	anything	401	148	himself
302	191	black	352	166	university	402	148	negative
303	191	schools	353	165	set	403	147	bears
304	190	actually	354	165	simply	404	147	matter
305	190	cannot	355	164	countries	405	147	middle
306	189	game	356	164	heart	406	146	became
307	189	throughout	357	164	student	407	146	getting
308	188	girl	358	163	cause	408	146	sense
309	188	wanted	359	163	political	409	146	taking
310	187	means	360	162	left	410	145	experience
311	186	little	361	162	lincoln	411	145	purpose
312	186	question	362	162	type	412	145	under
313	185	culture	363	162	white	413	15	writing
314	185	number	364	161	century	414	144	especially
315	184	comes	365	161	create	415	144	saying
316	184	pc	366	161	sports	416	143	along
317	184	support	367	160	rather	417	143	four
318	183	personal	368	160	view	418	143	likely
319	183	seen	369	159	looking	419	143	yet
320	183	within	370	159	positive	420	142	try
321	182	ever	371	158	animals	421	141	big
322	182	major	372	157	large	422	141	georgia
323	182	power	373	156	story	423	141	knowledge
324	181	candide	374	155	bush	424	141	similar
325	181	done	375	155	possible	425	140	group
326	181	ideas	376	155	research	426	140	plan
327	180	advertisement	377	155	troops	427	139	economy
328	180	hard	378	155	until	428	139	medium
329	180	taken	379	154	kids	429	139	water
330	180	u	380	154	last	430	139	whole
331	179	almost	381	154	might	431	138	author
332	178	case	382	154	rights	432	138	jobs
333	178	real	383	154	wants	433	138	modern
334	178	web	384	153	name	434	138	popular
335	177	games	385	153	needs	435	138	rasselas
336	177	speech	386	152	common	436	137	athletes
337	177	therefore	387	152	increase	437	137	five
338	174	amount	388	152	job	438	137	paper
339	174	changed	389	152	seem	439	137	sexual
340	173	future	390	151	learn	440	136	called
341	173	nothing	391	151	timothy	441	136	far
342	172	created	392	150	issues	442	136	goes
343	170	based	393	150	main	443	135	soldiers
344	170	living	394	150	mind	444	134	read
345	170	understand	395	150	next	445	134	reader
346	169	article	396	149	changes	446	134	relationship
347	169	control	397	149	individual	447	134	turn
348	169	hand	398	149	situation	448	134	word
349	168	nature	399	148	act	449	134	written
350	168	picture	400	148	already	450	133	free

451	133	result	501	118	stated	551	109	effects
452	133	self	502	117	began	552	109	face
453	132	again	503	117	clinton	553	109	male
454	132	realize	504	117	content	554	109	safe
455	132	reasons	505	117	involved	555	109	viewer
456	131	days	506	117	lead	556	108	easily
457	131	national	507	117	took	557	108	laws
458	131	product	508	116	business	558	108	playing
459	130	allowed	509	116	company	559	108	reality
460	130	freedom	510	116	longer	560	108	renaissance
461	130	marriage	511	116	side	561	107	background
462	130	started	512	115	ability	562	107	choose
463	129	came	513	115	actions	563	107	else
464	129	later	514	115	allow	564	107	itself
465	129	role	515	115	easy	565	107	opinion
466	129	small	516	115	individuals	566	107	presidential
467	129	wrong	517	115	looks	567	107	types
468	128	completely	518	115	mother	568	106	king
469	128	law	519	115	reading	569	106	players
470	128	several	520	115	towards	570	106	user
471	127	happiness	521	114	answer	571	106	visual
472	127	study	522	114	gives	572	106	went
473	127	teachers	523	114	needed	573	105	chris
474	127	users	524	114	protect	574	105	father
475	126	behind	525	114	second	575	105	fight
476	126	computers	526	114	sites	576	105	hours
477	126	learning	527	114	sure	577	105	images
478	126	shown	528	113	bad	578	105	physical
479	126	uses	529	113	becoming	579	105	police
480	125	lost	530	113	decision	580	105	science
481	125	questions	531	113	female	581	105	sometimes
482	125	windows	532	113	provide	582	104	ads
483	124	mass	533	113	strong	583	104	appearance
484	123	argument	534	113	usually	584	104	douglas
485	123	style	535	112	aspects	585	104	upon
486	122	continue	536	112	environment	586	103	esperanza
487	122	everyday	537	112	probably	587	103	lack
488	122	present	538	112	simple	588	103	success
489	121	book	539	112	tell	589	102	among
490	121	online	540	112	text	590	102	clear
491	121	violent	541	112	views	591	102	coach
492	120	community	542	111	allows	592	102	outside
493	120	head	543	111	effective	593	102	th
494	120	here	544	111	generation	594	101	bags
495	119	appeal	545	111	percent	595	101	beginning
496	119	kind	546	111	religion	596	101	considered
497	119	process	547	110	across	597	101	open
498	119	treadwell	548	110	companies	598	101	working
499	119	weapons	549	110	either	599	100	caused
500	118	iraqi	550	110	energy	600	100	mcluhan

APPENDIX B
Right & Left Collocates of *To*
(With prepositional *to* highlighted)

1	1462	to the	51	41	to realize	101	26	to hold
2	1203	to be	52	39	to fight	102	26	to improve
3	308	to a	53	39	to provide	103	26	to run
4	290	to make	54	39	to tell	104	26	to women
5	275	to do	55	38	to all	105	25	to continue
6	247	to have	56	37	to gain	106	25	to ensure
7	237	to get	57	37	to him	107	25	to establish
8	166	to see	58	37	to it	108	25	to survive
9	155	to help	59	37	to pay	109	25	to you
10	153	to find	60	36	to our	110	24	to any
11	145	to take	61	36	to read	111	24	to ask
12	135	to use	62	36	to write	112	24	to better
13	134	to their	63	35	to stop	113	24	to determine
14	131	to go	64	35	to these	114	24	to that
15	123	to keep	65	34	to leave	115	24	to wear
16	103	to say	66	34	to produce	116	23	to communicate
17	98	to become	67	34	to talk	117	23	to maintain
18	98	to live	68	33	to one	118	23	to spend
19	90	to her	69	33	to people	119	23	to your
20	89	to show	70	32	to choose	120	22	to accept
21	86	to give	71	32	to eat	121	22	to blame
22	81	to protect	72	32	to follow	122	22	to call
23	81	to this	73	32	to start	123	22	to connect
24	80	to them	74	31	to appeal	124	22	to lose
25	79	to his	75	31	to many	125	22	to meet
26	77	to create	76	31	to school	126	21	to allow
27	76	to change	77	30	to answer	127	21	to break
28	76	to know	78	30	to feel	128	21	to develop
29	73	to work	79	30	to increase	129	21	to fix
30	71	to an	80	30	to reach	130	21	to relate
31	71	to look	81	30	to sell	131	21	to want
32	68	to learn	82	29	to deal	132	20	to America
33	65	to play	83	29	to end	133	20	to and
34	62	to think	84	29	to how	134	20	to explain
35	59	to come	85	29	to prove	135	20	to happen
36	57	to try	86	29	to save	136	20	to its
37	55	to stay	87	29	to watch	137	20	to move
38	52	to what	88	28	to another	138	20	to pass
39	51	to believe	89	28	to attend	139	20	to present
40	51	to understand	90	28	to decide	140	20	to purchase
41	50	to my	91	28	to express	141	20	to teach
42	48	to me	92	28	to share	142	19	to act
43	48	to not	93	28	to war	143	19	to add
44	46	to prevent	94	27	to achieve	144	19	to as
45	46	to speak	95	27	to bring	145	19	to avoid
46	42	to other	96	27	to build	146	19	to begin
47	42	to support	97	27	to hear	147	19	to focus
48	42	to worry	98	27	to vote	148	19	to identify
49	41	to buy	99	26	to further	149	19	to obtain
50	41	to put	100	26	to grow	150	19	to receive

151	19	to send	193	14	to defend	235	12	to seek
152	19	to solve	194	14	to discover	236	12	to set
153	19	to those	195	14	to every	237	12	to students
154	19	to turn	196	14	to just	238	12	to which
155	19	to us	197	14	to kill	239	11	to actually
156	18	to control	198	14	to life	240	11	to American
157	18	to death	199	14	to more	241	11	to analyze
158	18	to describe	200	14	to persuade	242	11	to compare
159	18	to everyone	201	14	to pick	243	11	to cover
160	18	to Iraq	202	14	to pursue	244	11	to decrease
161	18	to let	203	14	to recognize	245	11	to discuss
162	18	to perform	204	14	to represent	246	11	to explore
163	18	to promote	205	14	to respond	247	11	to face
164	18	to view	206	14	to serve	248	11	to having
165	17	to being	207	14	to The	249	11	to love
166	17	to capture	208	13	to access	250	11	to raise
167	17	to figure	209	13	to accomplish	251	11	to return
168	17	to marry	210	13	to carry	252	11	to rise
169	17	to men	211	13	to cause	253	11	to society
170	17	to mind	212	13	to each	254	11	to some
171	17	to remain	213	13	to escape	255	10	to affect
172	17	to someone	214	13	to even	256	10	to apply
173	17	to stand	215	13	to feed	257	10	to class
174	16	to address	216	13	to hide	258	10	to convince
175	16	to draw	217	13	to interact	259	10	to different
176	16	to enter	218	13	to invade	260	10	to engage
177	16	to really	219	13	to join	261	10	to form
178	16	to reduce	220	13	to occur	262	10	to happiness
179	16	to succeed	221	13	to offer	263	10	to impress
180	15	to attract	222	13	to spread	264	10	to music
181	15	to experience	223	13	to suffer	265	10	to pull
182	15	to fit	224	12	to catch	266	10	to remember
183	15	to listen	225	12	to die	267	10	to replace
184	15	to others	226	12	to drink	268	10	to search
185	15	to such	227	12	to enjoy	269	10	to sit
186	15	to treat	228	12	to fulfill	270	10	to state
187	15	to walk	229	12	to inform	271	10	to study
188	15	to why	230	12	to lead	272	10	To the
189	15	to win	231	12	to only (50%)	273	10	to themselves
190	14	to college	232	12	to participate	274	10	to withdraw
191	14	to commit	233	12	to portray	275	9	to appear
192	14	to convey	234	12	to remove	276	9	to assist

Nominal Left Colligates of preposition *to* occurring more than once in 500K words

1476x	**…to the**	2x	day to the	**52x**	**…to what**
17x	solution(s) to the		end (n.) to the	2x	limit (n.) to what
			exception to the		
16x	attention to the		eyes (n.) to the	**48x**	**…to me**
			factor (n.) to the	2x	sense to me
9x	access (n.) to the		game to the		
	response to the		insight to the	**42x**	**…to other**
			life to the	2x	comparison to other
8x	way(s) to the		link (n.) to the		
			money to the	**39x**	**…to all**
7x	answer(s) (n.) to the		music to the	3x	answer(s) (n.) to all
	threat to the		people to the		
			resemblance to the	**37x**	**…to it**
6x	appeal(ing) (n.) to the		shock (n.) to the	3x	look (n.) to it
	contrast (n.) to the		sides to the		
	comparison to the		sites to the	**33x**	**…to one**
	message to the		speech to the	2x	billion to one
			statistics to the		culture to one
5x	addition to the		stop (n.) to the		
	improvement(s) to the		team to the	**28x**	**…to another**
	regard(s) (n.) to the		testament to the	2x	culture to another
			times to the		
4x	connection to the		transportation to the	**23x**	**…to your**
	contribution to the		trips to the	2x	access (n.) to your
	opposition to the		value (n.) to the		candidate to your
			viewer to the		response to your
3x	aspect(s) to the				
	audience to the	**308x**	**…to a/n**	**18x**	**…to Iraq**
	cost(s) (n.) to the	6x	right to a/n	2x	military to Iraq
	harm (n.) to the	5x	way to a/n		
	injury to the	4x	birth to a	**17x**	**…to men**
	key to the	3x	access (n.) to a	2x	advice to men
	knowledge to the		thanks to a/n		
	movement to the	2x	day to a	**14x**	**…to life**
	part to the		key to a	2x	right to life
	relation to the		user to a		
	thanks to the			**14x**	**…to more**
	victim to the	**134x**	**…to their**	2x	message to more
		2x	access (n.) to their		
2x	aid (n.) to the		attention to their	**12x**	**…to which**
	alternative to the		respect (n.) to their	2x	note (n.) to which
	approach (n.) to the				
	benefit (n.) to the	**81x**	**…to this**		
	blood to the	6x	solution to this		
	bonus to the	4x	answer (n.) to this		
	century to the	3x	addition to this		
	damage (n.) to the	2x	key to this		

TOTALS: Nominal Left Colligates of preposition *to*

39x access to
solution(s) to

30x attention to

27x response(s) to

26x addition to

23x answer(s) to
way(s) to

19x regard(s) to

16x key(s) to

14x threat(s) to

13x end to
right to

12x message(s) to

11x alternative(s) to
connection(s) to
life to
relation to
speech(es) to
thanks to

10x benefit(s) to
day to

9x birth to
comparison to
culture(s) to

8x contrast to
eye(s) to

7x approach(es) to
aspect(s) to
game(s) to
improvement(s) to
opposition to
way to

6x contribution(s) to
insight to
victim(s) to

5x advice to
audience to
injury to
money to
part to
respect to
side(s) to
trip(s) to
value to

4x billion to
damage to
candidate to
harm to
limit to
movement to
team to
thanks to
transportation to

3x exception to
factor(s) to
music to
shock to
stop to
statistics to

2x resemblance to
testament to

APPENDIX C
Left Collocates of *Of*
(With nominal left colligates highlighted)

1	313	one of	51	40	much of	101	24	benefits of
2	240	because of	52	40	outside of	102	24	case of
3	231	out of	53	39	side of	103	24	goal of
4	191	part of	54	38	All of	104	24	ideas of
5	189	all of	55	38	effects of	105	24	Some of
6	177	use of	56	37	hundreds of	106	23	attention of
7	153	amount of	57	37	States of	107	23	definition of
8	134	type of	58	37	years of	108	23	freedom of
9	124	number of	59	36	forms of	109	23	full of
10	122	lot of	60	36	issue of	110	23	future of
11	109	instead of	61	36	parts of	111	23	images of
12	99	some of	62	36	point of	112	23	list of
13	96	many of	63	36	variety of	113	23	middle of
14	94	form of	64	35	life of	114	23	story of
15	92	idea of	65	35	sort of	115	23	those of
16	91	One of	66	35	source of	116	22	fear of
17	90	types of	67	35	view of	117	22	generation of
18	89	kind of	68	34	history of	118	22	knowledge of
19	88	way of	69	34	image of	119	22	meaning of
20	84	most of	70	34	piece of	120	22	picture of
21	84	people of	71	33	appearance of	121	22	pictures of
22	81	lack of	72	32	aware of	122	22	population of
23	78	sense of	73	32	top of	123	22	production of
24	74	majority of	74	32	world of	124	21	ahead of
25	71	aspects of	75	31	period of	125	21	any of
26	67	result of	76	30	care of (v.)	126	21	death of
27	67	University of	77	30	development of	127	21	hopes of
28	67	weapons of	78	30	each of	128	21	method of
29	64	end of	79	30	Most of	129	21	nature of
30	62	aspect of	80	30	style of	130	20	capable of
31	61	lives of	81	30	thought of	131	20	couple of
32	61	that of	82	30	understanding of	132	20	level of
33	60	percent of	83	29	cause of	133	20	problem of
34	60	thousands of	84	29	control of	134	20	role of
35	59	group of	85	29	loss of	135	19	both of
36	58	rest of	86	29	quality of	136	19	characteristics of
37	56	example of	87	29	risk of	137	19	combination of
38	56	more of	88	28	terms of	138	19	content of
39	51	age of	89	27	amounts of	139	19	cost of
40	48	millions of	90	27	citizens of	140	19	evidence of
41	47	purpose of	91	27	course of	141	19	is of
42	45	time of	92	27	means of	142	19	name of
43	44	importance of	93	27	off of	143	19	presence of
44	44	Instead of	94	26	chance of	144	18	advantage of
45	44	state of	95	26	hours of	145	18	cases of
46	43	Because of	96	26	Many of	146	18	choice of
47	43	beginning of	97	25	center of	147	18	city of
48	43	examples of	98	25	creation of	148	18	Declaration of
49	43	front of	99	25	half of	149	18	kinds of
50	41	think of	100	25	process of	150	18	matter of

151	18	message of
152	18	plenty of
153	18	set of
154	18	ways of
155	18	women of
156	17	culture of
157	17	game of
158	17	members of
159	17	question of
160	17	regardless of
161	17	version of
162	16	color of
163	16	concept of
164	16	custody of
165	16	days of
166	16	favor of
167	16	leader of
168	16	member of
169	16	numbers of
170	16	pair of
171	16	perception of
172	16	pounds of
173	16	rise of
174	16	size of
175	16	support of
176	15	actions of
177	15	author of
178	15	background of
179	15	bit of
180	15	charge of
181	15	division of
182	15	eyes of
183	15	groups of
184	15	job of
185	15	medium of
186	15	up of
187	14	consists of
188	14	emotions of
189	14	feeling of
190	14	feelings of
191	14	help of
192	14	methods of
193	14	none of
194	14	occupation of
195	14	opinion of
196	14	percentage of
197	14	President of
198	14	president of
199	14	success of
200	14	system of
201	14	third of
202	14	treatment of
203	14	two of
204	14	words of
205	13	act of
206	13	areas of
207	13	country of
208	13	dangers of
209	13	effect of
210	13	elements of
211	13	heart of
212	13	need of
213	13	possibility of
214	13	rate of
215	13	reality of
216	13	rid of
217	13	rights of
218	13	safety of
219	13	span of
220	13	stage of
221	13	structure of
222	13	theme of
223	12	accused of
224	12	acts of
225	12	apart of [*sic*]
226	12	body of
227	12	chances of
228	12	class of
229	12	consisted of
230	12	cover of
231	12	enough of
232	12	establishment of
233	12	hands of
234	12	land of
235	12	place of
236	12	principles of
237	12	product of
238	12	Regardless of
239	12	sight of
240	12	times of
241	12	title of
242	12	works of
243	12	year of
244	11	and of
245	11	are of
246	11	As of
247	11	back of
248	11	collection of
249	11	consequences of
250	11	day of
251	11	debate of
252	11	debates of
253	11	Department of
254	11	description of
255	11	director of
256	11	fun of
257	11	History of
258	11	inside of
259	11	introduction of
260	11	invention of
261	11	levels of
262	11	look of
263	11	metaphor of
264	11	outcome of
265	11	pieces of
266	11	power of
267	11	problems of
268	11	range of
269	11	representation of
270	11	signs of
271	11	taste of
272	11	topic of
273	11	Vision of
274	11	word of
275	10	afraid of
276	10	area of
277	10	as of
278	10	billions of
279	10	Both of
280	10	children of
281	10	component of
282	10	deaths of
283	10	decision of
284	10	face of
285	10	few of
286	10	growth of
287	10	hiring of
288	10	influence of
289	10	length of
290	10	man of
291	10	me of
292	10	points of
293	10	portion of
294	10	quarter of
295	10	responsibility of
296	10	results of
297	10	section of
298	10	separation of
299	10	series of
300	10	sign of

301	10	spread of	351	8	mode of	401	7	setup of
302	10	start of	352	8	option of	402	7	smell of
303	10	theory of	353	8	parents of	403	7	speed of
304	10	total of	354	8	perspective of	404	7	students of
305	10	value of	355	8	plethora of	405	7	thoughts of
306	9	absence of	356	8	price of	406	7	tons of
307	9	Act of	357	8	principle of	407	7	views of
308	9	best of	358	8	purposes of	408	7	violation of
309	9	character of	359	8	share of	409	7	work of
310	9	components of	360	8	son of	410	6	acceptance of
311	9	danger of	361	8	sorts of	411	6	argument of
312	9	deal of	362	8	stereotypes of	412	6	basis of
313	9	destruction of	363	8	subject of	413	6	behavior of
314	9	ethics of	364	8	them of	414	6	Bill of
315	9	extension of	365	8	unaware of	415	6	bottle of
316	9	fall of	366	8	usage of	416	6	bunch of
317	9	First of	367	8	users of	417	6	causes of
318	9	foundation of	368	8	waste of	418	6	concern of
319	9	free of	369	7	awareness of	419	6	conclusion of
320	9	health of	370	7	being of	420	6	conditions of
321	9	minds of	371	7	birth of	421	6	Defense of
322	9	needs of	372	7	colors of	422	6	desire of
323	9	opinions of	373	7	consumption of	423	6	Effects of
324	9	Out of	374	7	corner of	424	6	element of
325	9	sides of	375	7	cup of	425	6	execution of
326	9	sources of	376	7	degree of	426	6	family of
327	9	stories of	377	7	details of	427	6	format of
328	9	styles of	378	7	direction of	428	6	four of
329	9	supply of	379	7	economy of	429	6	funding of
330	9	war of	380	7	experience of	430	6	genre of
331	8	advancement of	381	7	experiences of	431	6	good of
332	8	ages of	382	7	expression of	432	6	government of
333	8	analysis of	383	7	fact of	433	6	him of
334	8	array of	384	7	fundamentals of	434	6	House of
335	8	beliefs of	385	7	guilty of	435	6	increase of
336	8	bottom of	386	7	invasion of	436	6	interpretation of
337	8	change of	387	7	line of	437	6	issues of
338	8	community of	388	7	mind of	438	6	Journal of
339	8	consist of	389	7	mixture of	439	6	lots of
340	8	Each of	390	7	notion of	440	6	midst of
341	8	equality of	391	7	opportunity of	441	6	mother of
342	8	existence of	392	7	order of	442	6	movement of
343	8	feature of	393	7	Part of	443	6	nation of
344	8	flow of	394	7	portrayal of	444	6	out-of
345	8	focus of	395	7	position of	445	6	People of
346	8	institution of	396	7	presentation of	446	6	photo of
347	8	logos of	397	7	protection of	447	6	portions of
348	8	made of	398	7	pursuit of	448	6	pound of
349	8	memories of	399	7	sake of	449	6	regulation of
350	8	men of	400	7	search of	450	6	removal of

APPENDIX D
Left Collocates of *In*
(With nominal left colligates highlighted)

1	100	is in	51	21	work in (v.)	101	13	successful in
2	98	people in	52	20	stated in	102	13	this in
3	92	and in	53	20	than in	103	13	troops in
4	82	change in	54	19	difference in	104	12	debate in
5	79	are in	55	19	keep in	105	12	differences in
6	67	war in	56	19	men in	106	12	fit in
7	64	involved in	57	19	only in	107	12	here in
8	59	live in	58	19	ways in	108	12	located in
9	58	be in	59	18	especially in	109	12	occur in
10	58	women in	60	17	do in	110	12	presence in
11	56	up in	61	17	issue in	111	12	succeed in
12	52	but in	62	17	participate in	112	12	taken in
13	51	role in	63	17	problems in	113	12	violence in
14	49	increase in	64	17	resulted in	114	12	when in
15	47	them in	65	17	situation in	115	11	aid in
16	47	was in	66	17	still in	116	11	all in
17	44	used in	67	16	country in	117	11	also in
18	43	living in (v.)	68	16	day in	118	11	article in
19	37	it in	69	16	everything in	119	11	did in
20	37	War in	70	16	made in	120	11	displayed in
21	35	out in	71	16	results in (v.)	121	11	education in
22	35	that in	72	15	being in	122	11	herself in
23	34	changes in	73	15	engage in	123	11	lies in (v.)
24	32	interest in	74	15	even in	124	11	more in
25	32	interested in	75	15	factor in	125	11	portrayed in
26	32	place in	76	15	for in	126	11	power in
27	31	found in	77	15	have in	127	11	present in
28	31	point in	78	15	problem in	128	11	remain in
29	30	time in	79	15	set in (v.)	129	11	step in (v.)
30	29	believe in	80	15	so in	130	11	taught in
31	29	seen in	81	15	which in	131	11	with in
32	29	stay in	82	14	began in	132	11	world in
33	26	on in	83	14	believed in	133	10	candidate in
34	26	placed in	84	14	characters in	134	10	child in
35	26	were in	85	14	Iraq in	135	10	done in
36	25	put in	86	14	one in	136	10	everyone in
37	25	result in (v.)	87	14	resulting in	137	10	females in
38	24	children in	88	14	technology in	138	10	fought in
39	24	life in	89	14	way in	139	10	good in (adj.)
40	24	themselves in	90	13	advances in	140	10	happiness in
41	23	lived in	91	13	as in	141	10	him in
42	23	lives in (v.)	92	13	audience in	142	10	like in
43	22	been in	93	13	dressed in	143	10	now in
44	22	part in	94	13	form in	144	10	play in (v.)
45	22	shown in	95	13	girl in	145	10	presented in
46	22	things in	96	13	However, in	146	10	prevalent in
47	22	written in	97	13	information in	147	10	rise in
48	21	not in	98	13	published in	148	10	school in
49	21	or in	99	13	raised in	149	10	schools in
50	21	students in	100	13	sitting in	150	10	something in

APPENDIX E
Left Collocates of *For*
(With nominal left colligates highlighted)

1	49	reason for
2	45	responsible for
3	37	looking for
4	36	known for
5	34	need for
6	32	is for
7	32	up for
8	29	and for
9	29	order for
10	28	reasons for
11	28	search for
12	28	time for
13	27	fighting for
14	26	used for
15	23	fight for (v.)
16	23	room for
17	20	but for
18	20	easier for
19	20	not for
20	19	life for
21	19	out for
22	18	best for
23	18	difficult for
24	18	it for
25	17	look for (v.)
26	17	plan for (v.)
27	17	respect for
28	17	vote for (v.)
29	16	change for
30	16	money for
31	16	pay for (v.)
32	16	was for
33	16	way for
34	16	work for (v.)
35	15	allow for
36	15	hard for
37	15	place for
38	14	blamed for
39	14	candidate for
40	14	good for
41	14	possible for
42	13	allows for
43	13	are for
44	13	enough for
45	12	Iraq for
46	12	love for
47	12	name for
48	12	responsibility for
49	12	support for
50	12	them for
51	11	demand for
52	11	play for (v.)
53	11	provide for
54	11	stands for (v.)
55	11	there for
56	10	be for
57	10	concern for
58	10	except for
59	10	him for
60	10	impossible for
61	10	made for
62	10	metaphor for
63	10	necessary for
64	10	or for
65	10	prepared for
66	10	searching for
67	10	stand for (v.)
68	9	allowed for
69	9	basis for
70	9	blame for
71	9	care for (v.)
72	9	cause for
73	9	different for
74	9	hope for
75	9	important for
76	9	just for
77	9	on for
78	9	opportunity for
79	9	Paws for
80	8	arrested for
81	8	available for
82	8	candidates for
83	8	easy for
84	8	education for
85	8	lives for (v.)
86	8	opportunities for
87	8	running for
88	8	solution for
89	8	standard for
90	8	strive for
91	8	want for
92	8	Watts for
93	7	ad for
94	7	advertisement for
95	7	around for
96	7	As for
97	7	country for
98	7	coverage for
99	7	created for
100	7	debates for

APPENDIX F
Left Collocates of *With*
(With nominal left colligates highlighted)

1	82	along with	42	11	happy with	83	6	comply with
2	81	up with	43	11	them with	84	6	experience with
3	52	deal with (v.)	44	11	women with	85	6	interview with
4	47	agree with	45	10	child with	86	6	issue with
5	46	associated with	46	10	compete with	87	6	men with
6	41	do with	47	10	friends with	88	6	replaced with
7	38	people with	48	10	home with	89	6	trade with
8	32	problem with	49	10	work with(v.)	90	6	trouble with
9	30	and with	50	9	come with	91	6	works with
10	29	deals with(v.)	51	9	done with	92	5	accordance with
11	28	Along with	52	9	interfere with	93	5	agrees with
12	28	dealing with	53	9	living with	94	5	around with
13	26	touch with	54	9	love with	95	5	associate with
14	24	out with	55	9	off with	96	5	athletes with
15	23	filled with	56	9	playing with	97	5	better with
16	21	disagree with	57	9	relationships with	98	5	case with
17	20	interact with	58	9	sex with	99	5	computer with
18	19	communicate with	59	9	working with	100	5	content with
19	19	dealt with	60	9	world with	101	5	coupled with
20	19	relationship with	61	9	wrong with	102	5	debate with
21	19	war with	62	8	be with	103	5	equipped with
22	18	concerned with	63	8	Even with	104	5	fascination with
23	16	away with	64	8	help with (v.)	105	5	him with
24	16	identify with	65	8	it with	106	5	in with
25	16	play with	66	8	life with	107	5	individuals with
26	15	but with	67	8	man with	108	5	infected with
27	15	familiar with	68	8	one with	109	5	interacting with
28	15	involved with	69	8	satisfied with			
29	15	time with	70	7	children with	110	5	job with
30	13	contact with	71	7	comfortable with	111	5	made with
31	13	live with	72	7	connection with	112	5	not with
32	13	problems with	73	7	information with	113	5	obsessed with
33	13	that with	74	7	interaction with	114	5	on with
34	12	connect with	75	7	is with	115	5	reader with
35	12	faced with	76	7	issues with	116	5	room with
36	12	struggle with (v.)	77	7	left with	117	5	so with
37	12	themselves with	78	7	met with	118	5	struggling with
38	11	begin with	79	7	someone with			
39	11	charged with	80	6	ad with	119	5	through with
40	11	comes with	81	6	are with	120	5	viewer with
41	11	diagnosed with	82	6	begins with			

APPENDIX G
Left Collocates of *On*
(With nominal left colligates highlighted)

1	92	based on	45	11	opinions on	89	6	audience on
2	44	going on	46	11	War on	90	6	blamed on
3	41	focus on (v.)	47	10	back on	91	6	debates on
4	41	war on	48	10	but on	92	6	decide on
5	40	goes on	49	10	depends on	93	6	found on
6	36	impact on	50	10	lives on (v.)	94	6	her on
7	30	effect on	51	10	money on	95	6	knowledge on
8	27	focused on	52	10	music on	96	6	pressure on
9	26	out on	53	10	solely on	97	6	relying on
10	26	views on	54	10	went on	98	6	this on
11	26	was on	55	9	been on	99	6	work on
12	25	is on	56	9	being on	100	5	affect on
13	24	information on	57	9	depend on	101	5	attack on
14	24	up on	58	9	dependent on	102	5	Based on
15	23	and on	59	9	emphasis on	103	5	child on
16	23	are on	60	9	hands on	104	5	children on
17	23	put on	61	9	life on	105	5	concentrate on
18	23	rely on	62	9	look on (v.)	106	5	decisions on
19	20	them on	63	9	made on	107	5	done on
20	19	depending on	64	9	things on	108	5	fly on (v.)
21	19	effects on	65	9	view on	109	5	food on
22	18	go on	66	8	better on	110	5	founded on
23	17	House on	67	8	debate on	111	5	get on
24	17	working on	68	8	had on	112	5	girl on
25	16	be on	69	8	has on	113	5	him on
26	16	more on	70	8	influence on	114	5	hold on
27	16	spent on	71	8	later on	115	5	hours on
28	15	have on	72	8	relies on	116	5	Later on
29	15	in on	73	8	see on	117	5	man on
30	15	people on	74	8	sitting on	118	5	off on
31	14	down on	75	8	stand on	119	5	outlook on
32	14	placed on	76	8	were on	120	5	perspective on
33	14	stance on	77	7	icons on	121	5	pictures on
34	14	women on	78	7	move on	122	5	place on
35	13	it on	79	7	or on	123	5	research on
36	13	seen on	80	7	person on	124	5	restrictions on
37	13	time on	81	7	so on	125	5	set on
38	12	focusing on	82	7	than on	126	5	spend on
39	12	heavily on	83	7	words on	127	5	takes on
40	12	opinion on	84	7	written on	128	5	that on
41	12	take on	85	6	an on	129	5	toll on
42	11	focuses on	86	6	appear on	130	5	use on (v.)
43	11	live on	87	6	attacks on	131	5	used on
44	11	not on	88	6	attention on	132	5	voted on

APPENDIX H
Left Collocates of *By*
(With nominal left colligates highlighted)

1	30	caused by	29	7	worn by
2	26	affected by	30	6	abused by
3	26	written by	31	6	article by
4	18	and by	32	6	controlled by
5	18	made by	33	6	defined by
6	16	done by	34	6	judged by
7	16	influenced by	35	6	killed by
8	15	is by	36	6	known by
9	15	used by	37	6	live by
10	12	followed by	38	6	not by
11	12	that by	39	6	or by
12	11	this by	40	6	out by
13	10	but by	41	6	passed by
14	10	created by	42	6	simply by
15	10	them by	43	6	told by
16	9	conducted by	44	5	approached by
17	9	supported by	45	5	are by
18	9	up by	46	5	asked by
19	8	accompanied by	47	5	audience by
20	8	given by	48	5	dominated by
21	8	her by	49	5	increased by
22	8	provided by	50	5	life by
23	8	shown by	51	5	off by
24	7	abide by	52	5	overwhelmed by
25	7	get by	53	5	presented by
26	7	held by	54	5	protected by
27	7	produced by	55	5	run by
28	7	surrounded by	56	5	seen by

APPENDIX I
Left Collocates of *From*
(With nominal left colligates highlighted)

1	87	away from
2	35	different from
3	33	comes from
4	30	come from
5	23	people from
6	21	suffer from
7	19	them from
8	16	learn from
9	12	and from
10	12	came from
11	11	benefit from (v.)
12	11	it from
13	11	suffering from
14	10	everything from
15	10	far from
16	10	range from
17	9	attention from
18	9	changed from
19	9	coming from
20	9	died from
21	9	is from
22	9	removed from
23	9	themselves from
24	9	troops from
25	9	us from
26	8	children from
27	8	resulting from
28	8	water from
29	7	anything from
30	7	apart from
31	7	free from
32	7	ranging from
33	7	stems from (v.)
34	7	support from
35	7	up from
36	7	withdraw from
37	6	back from (adv.)
38	6	cells from
39	6	derived from
40	6	escape from
41	6	gone from
42	6	graduated from
43	6	her from
44	6	home from (adv.)
45	6	information from
46	6	lot from
47	6	protected from
48	6	States from
49	5	Aside from
50	5	citizens from
51	5	example from
52	5	freedom from
53	5	girl from
54	5	goods from
55	5	graduate from (v.)
56	5	him from
57	5	kids from
58	5	money from
59	5	moved from
60	5	right from (adv.
61	5	suffered from
62	5	taken from
63	5	this from
64	5	transition from
65	5	women from
66	4	advice from

APPENDIX J
Left Collocates of *At*
(With nominal left colligates highlighted)

1	75	look at (v.)
2	61	looking at
3	24	are at
4	21	is at
5	21	looked at
6	20	and at
7	16	or at
8	15	was at
9	14	be at
10	12	looks at (v.)
11	12	people at
12	10	were at
13	9	but at
14	8	here at
15	8	up at
16	7	all at
17	7	him at
18	7	Looking at
19	7	not at
20	7	out at
21	7	present at
22	6	for at
23	6	issue at
24	6	it at
25	5	around at
26	5	arrived at
27	5	back at
28	5	bags at
29	5	chance at
30	5	fetus at
31	5	food at
32	5	good at
33	5	life at
34	5	students at
35	5	that at
36	5	war at

APPENDIX K
Left Collocates of *About*
(With nominal left colligates highlighted)

1	45	worry about
2	41	think about
3	32	information about
4	31	talking about (v.)
5	30	talk about (v.)
6	29	is about
7	27	more about
8	21	care about (v.)
9	19	talks about (v.)
10	18	all about
11	18	know about
12	16	talked about
13	14	was about
14	13	brought about
15	13	something about
16	12	much about
17	12	things about
18	11	write about
19	10	say about
20	10	thinking about (v.)
21	9	anything about
22	9	concerned about
23	9	feel about
24	9	for about
25	9	heard about
26	9	just about
27	9	out about
28	9	worried about
29	8	bring about
30	8	complain about
31	8	knowledge about
32	8	Think about
33	8	thought about (v.)
34	7	asked about
35	7	be about
36	7	excited about
37	7	forget about
38	7	learn about
39	7	learning about
40	7	lot about
41	7	read about
42	7	views about
43	7	what about
44	7	What about
45	6	are about
46	6	concerns about
47	6	nothing about
48	6	only about
49	6	passionate about
50	6	questions about
51	6	s about
52	6	story about
53	6	were about
54	6	worrying about
55	6	writes about
56	5	came about
57	5	cared about
58	5	concern about
59	5	done about
60	5	facts about
61	5	feels about
62	5	have about
63	5	hearing about
64	5	in about
65	5	not about
66	5	people about
67	5	thing about
68	4	article about

APPENDIX L
Two-Word Clusters with Prepositions

Most Frequent Two-Word Clusters with Prepositions in UGALECT (N-grams)

of the	*of a*	*of this*	*of his*	*of these*	*of which*
in the	*in a*				*in which*
to the	*to a*				
on the	*on a*				
for the	*for a*				
with the	*with a*				
as the	*as a*				
from the					
by the					
at the					
about the					
into the					
over the					

N + P Clusters

part(s) of	247x	*time(s) of*	58x
type(s)of	224x	*effect(s) of*	57x
use(s) of	185x	*purpose(s) of*	55x
amount(s) of	180x	*increase(es) in*	54x
number(s) of	140x	*war(s) on*	53x
aspect(s) of	136x	*millions of*	52x
form(s) of	130x	*effect(s) on*	50x
lot(s) of	128x	*year(s) of*	49x
change(s) in	121x	*side(s) of*	48x
idea(s) of	116x	*point(s) of*	46x
war(s) in	110x	*beginning(s) of*	45x
kind(s) of	108x	*history of*	45x
way(s) of	107x	*importance of*	45x
example(s) of	100x	*picture(s) of*	45x
lives/life of	99x	*piece(s) of*	45x
people in	99x	*problem(s) with*	45x
state(s) of	91x	*sort(s) of*	45x
people of	90x	*source(s) of*	44x
lack of	82x	*case(s) of*	43x
sense of	79x	*front of*	43x
reason(s) for	77x	*issue(s) of*	43x
result(s) of	77x	*view(s) of*	43x
majority of	76x	*people with*	42x
women/woman in	75x	*difference(s) between*	40x
group(s) of	74x	*time(s) in*	40x
*weapons o*f	68x	*hundreds of*	39x
University of	67x	*access to* (prep)	39x
end of	65x	*solution(s) to* (prep)	39x
age(s) of	64x	*style(s) of*	39x
thousands of	64x	*chance(s) of*	38x
role(s) in	61x	*impact on*	37x
image(s) of	60x	*variety/ies of*	37x
percent of	60x	*cause(s) of*	36x
rest of	58x	*child(ren) in*	36x

period(s) of	36x	*hope(s) of*	26x
place(s) in	36x	*hours of*	26x
method(s) of	35x	*word(s) of*	26x
appearance(s) of	35x	*answer(s) to* (prep)	25x
need for	35x	*freedom(s) of*	25x
risk(s) of	35x	*generation(s) of*	25x
view(s) on	35x	*list(s) of*	25x
act(s) of	34x	*opinion(s) of*	25x
interest(s) in	34x	*process of*	25x
life/lives in	34x	*benefits of*	24x
member(s) of	34x	*country/ies in*	24x
world of	34x	*future of*	24x
death(s) of	33x	*information on*	24x
information about	33x	*people from*	24x
point(s) in	33x	*way(s) to* (prep)	24x
top(s) of	33x	*women/woman of*	24x
way(s) in	33x	*area(s) of*	23x
control(s) of	32x	*attention of*	23x
difference(s) in	32x	*candidate(s) for*	23x
level(s) of	32x	*color(s) of*	23x
problem(s) in	32x	*cost(s) of*	23x
quality/ies of	32x	*debate(s) of*	23x
story/ies of	32x	*definition of*	23x
development of	31x	*middle of*	23x
problem(s) of	31x	*need(s) of*	23x
attention to (prep)	30x	*opinion(s) on*	23x
citizen(s) of	30x	*plan(s) for*	23x
term(s) of	30x	*population of*	23x
understanding of	30x	*room for*	23x
feeling(s) of	29x	*danger(s) of*	22x
loss of	29x	*knowledge of*	22x
order for	29x	*pound(s) of*	22x
life/lives for	28x	*production of*	22x
meaning(s) of	28x	*school(s) in*	22x
president of	28x	*situation(s) in*	22x
relationship(s) with	28x	*advantage(s) of*	21x
time for	28x	*city/ies of*	21x
center(s) of	27x	*content(s) of*	21x
course of	27x	*man/men of*	21x
day(s) of	27x	*name(s) of*	21x
goal(s) of	27x	*nature of*	21x
issue(s) in	27x	*sign(s) of*	21x
means of	27x	*war(s) with*	21x
role(s) of	27x	*character(s) in*	20x
student(s) in	27x	*couple of*	20x
thing(s) in	27x	*day(s) in*	20x
addition to (prep)	26x	*debate(s) in*	20x
creation of	26x	*declaration of*	20x
fear(s) of	26x	*evidence of*	20x
half of	26x	*leader(s) of*	20x

message(s) of	20x	*favor of*	16x
principle(s) of	20x	*heart(s) of*	16x
right(s) of	20x	*medium of*	16x
set(s) of	20x	*key(s) to* (prep)	16x
technology/ies in	20x	*mind(s) of*	16x
		money for	16x
action(s) of	19x	*pair of*	16x
author(s) of	19x	*place(s) for*	16x
background(s) of	19x	*portion(s) of*	16x
characteristics of	19x	*reality/ies of*	16x
choice(s) of	19x	*support of*	16x
combination of	19x	*system(s) of*	16x
component(s) of	19x	*theme(s) of*	16x
element(s) of	19x	*theory/ies of*	16x
presence of	19x		
question(s) of	19x	*body/ies of*	15x
stage(s) of	19x	*child(ren) of*	15x
third(s) of	19x	*class(es) of*	15x
work(s) of	19x	*consequence(s) of*	15x
		description(s) of	15x
candidate(s) in	18x	*face(s) of*	15x
culture of	18x	*factor in*	15x
eye(s) of	18x	*form(s) in*	15x
game(s) of	18x	*people on*	15x
matter of	18x	*percentage(s) of*	15x
perception(s) of	18x	*possibility/ies of*	15x
plenty of	18x	*product(s) of*	15x
power(s) of	18x	*responsibility/ies of*	15x
rate(s) of	18x	*time with*	15x
respect for	18x	*title(s) of*	15x
size(s) of	18x	*treatment of*	15x
version(s) of	18x	*women/woman with*	15x
way(s) for	18x		
		advances in	14x
bit(s) of	17x	*control over*	14x
charge(s) of	17x	*cover(s) of*	14x
child(ren) with	17x	*help of*	14x
concept(s) of	17x	*name(s) for*	14x
country/ies of	17x	*occupation of*	14x
job(s) of	17x	*place(s) of*	14x
metaphor(s) of	17x	*span(s) of*	14x
rise of	17x	*structure of*	14x
stance on	17x	*success of*	14x
thing(s) about	17x	*taste of*	14x
vision(s) of	17x	*threat(s) to* (prep)	14x
		topic(s) of	14x
audience(s) in	16x		
custody of	16x	*change(s) for*	13x
division of	16x	*concern(s) for*	13x
emotion(s) of	16x	*contact with*	13x

billion(s) of	13x	*section(s) of*	12x
end to (prep)	13x	*sight of*	12x
establishment of	13x	*support for*	12x
hand(s) of	13x		
information in	13x	*article in*	11x
look(s) of	13x	*back of*	11x
metaphor(s) for	13x	*connection(s) to* (prep)	11x
people at	13x	*education in*	11x
relationship(s)between	13x	*friend(s) with*	11x
safety of	13x	*hands on*	11x
separation of	13x	*introduction of*	11x
system(s) in	13x	*length(s) of*	11x
troops in	13x	*power in*	11x
violence in	13x	*quarter(s) of*	11x
		range of	11x
change(s) to (prep)	12x	*rise(s) in*	11x
collection(s) of	12x	*series of*	11x
decision(s) of	12x	*value(s) of*	11x
demand(s) for	12x	*war against*	11x
Department of	12x	*world in*	11x
director of	12x		
female(s) in	12x	*connection between*	10x
influence(s) of	12x	*growth of*	10x
invention(s) of	12x	*happiness in*	10x
land of	12x	*hiring of*	10x
love for	12x	*money on*	10x
message(s) to (prep)	12x	*music on*	10x
outcome(s) of	12x	*spread of*	10x
presence in	12x	*start of*	10x
representation(s) of	12x	*total of*	10x
responsibility for	12x		

Phrasal Pronouns

one of	404x
all of	227x
some of	123x
many of	122x
most of	114x
more of	56x
much of	44x
each of	38x
both of	29x
one to (81% INF)	27x
any of	21x
everything in	21x
all in	16x
none of	16x
two of	16x
everyone in	15x
one in	15x
enough of	12x
everything from	11x
few of	10x
something in	10x

Phrasal Prepositions

because of	283x
out of	240x
according to	227x
due to	187x
instead of	153x
along with	110x
but for	22x
ahead of	21x
prior to	16x
next to	13x
inside of	11x
thanks to	11x
except for	10x

Three-word Prepositionals

in front of	37x
in order for	29x
in touch with	25x
in addition to	20x
in response to	20x
in favor of	13x
in contrast to	6x

Prepositional Phrases (aka Conjunctive Adverbials)

for example,	162x
in fact,	56x
of course,	51x
as a result,	50x
on the other hand,	49x
for instance,	42x
in my opinion,	41x
in conclusion,	33x
in addition,	23x
in other words,	13x
in the case of,	12x
on the contrary,	10x
in contrast,	8x