Concordia Working Papers in Applied Linguistics, 4, 2013

An Alternate Academic Vocabulary: A Word List for Canadian University Websites

Victoria Surtees
Concordia University

Marlise Horst
Concordia University

Abstract

The study of lexis in academic texts has been extensive. However, within this literature, little attention has been paid to a reading challenge that virtually all second language university students have in common: the information provided to international students via university websites on topics such as admission, immigration, and academic life. This study examined the lexical profile of a 147,000 word corpus composed of Canadian university website literature aimed at international students. The project investigated whether students with knowledge of the 2000 most frequent word families as identified in the British National Corpus could be reasonably expected to comprehend these sites, and if not, whether a list of field specific vocabulary could be created to enhance their comprehension. Results showed the 2000 most frequent families did not provide sufficient coverage of the corpus and that a further 226 lemmatized types provided the necessary coverage to attain the 95% target proposed as an appropriate figure for adequate comprehension (Nation, 2006). The characteristics of this 226 item list, referred to as the International Student Word List (ISWL), are discussed in reference to different approaches to word list creation for second language teaching.

To date, the study of lexis in academic texts has been extensive. Researchers have examined subtechnical vocabulary across disciplines, (e.g., Coxhead, 2000), compared the lexis of spoken and written university texts (e.g., Biber et al., 2002), and identified multiword strings appearing in academic literature (e.g., Biber, 2007; Liu, 2012). However, within these corpus studies, little attention has been paid to the literature that virtually all second language university students have in common: the information provided to international students via university websites. This literature addresses topics such as admission, program requirements, insurance, immigration, and academic life and may be as dense as university textbooks (Biber et al., 2002).

If the language presented on these websites is extremely specialized, it is possible that the vital information concerning international students' academic lives is not easily accessible. Can students with limited knowledge of English achieve adequate comprehension of these websites? Is there a specific list of vocabulary that, if known by students, might significantly increase their comprehension?

To investigate this issue, we will first review the literature on the creation of specialized word lists for L2 learning purposes. We will then report on the results of a study in which we examined the lexis in a corpus composed of electronic texts intended for international students attending Canadian universities.

TEXT COVERAGE AND WORD LISTS

Much of the impetus behind the creation of word lists is based on the link between reading comprehension and vocabulary knowledge. Studies have generally found that learners must know between 95-98% of words in a text to achieve adequate comprehension of its content (see Nation, 2001, 2006 for a review). Logically, this entails that by learning vocabulary that occurs frequently, students will be able to attain this 95% known-word threshold in less time than students taught vocabulary at random. Thus, by constructing word lists based on frequency, researchers are attempting to pre-select the vocabulary that provides the highest degree of text coverage (i.e., that account for a significant proportion of the vocabulary encountered in a given set of texts) and which, if learned, will allow students to progress more quickly (Coxhead, 2000).

Some word lists are based on the calculating the frequencies of *lemmas*, defined as a base word and its grammatically inflected forms. For example, the lemma for the verb *build*, includes *builds*, *built*, and *building*

but not *builder*. Another approach is to select items according to the frequencies of their *word families*. A word family includes basic derived forms such as *builder, buildings* and *rebuild* in addition to the inflected verb forms included in the concept of lemma. By adopting the word family approach, researchers assume that a learner who knows *build* will also recognize all its derived forms whereas a researcher adopting the lemma aproach assumes that the learner must be taught each lemma separately for adequate comprehension.

Items are selected for inclusion according to their frequency of occurrence as observed in a body of authentic texts, referred to as a corpus. Probably the most famous word list compiled in this way is West's General Service List (GSL, 1953) of the 2000 most frequent word families, created through the manual analysis of a five million word corpus of written texts. The word families on the GSL typically account for between 70-85% of the vocabulary in texts (Nation, 2001) and thus represent prime teaching targets. Since the creation of the GSL, corpora such as the 100 million-word *British National Corpus* (BNC) have gone digital, and are now analysed with powerful computer software, making word lists less time consuming to construct.

Specialized word lists, which are the focus of this paper, cater specifically to the needs of students in fields such as engineering, medicine, business or academia. The methods researchers use to create these lists vary according to the lists' intended purpose. The next sections will review the two predominant approaches that emerge in the literature.

The Layered Approach

The first approach, termed the "layered approach" (Coxhead & Hirsh, 2007), is intended for intermediate to advanced learners and assumes that the population for whom the list is created is already familiar with the 2000 most frequent words families. It aims to identify a manageable list beyond the first 2000 most frequent words that provides reasonable coverage of a specialized corpus, usually between 5% and 10%.

The layered approach was famously adopted by Coxhead (2000) to create the Academic Word List (AWL), a list of semi-specialized word families frequent in written academic texts. Hoping to improve upon the already existing University Word List of 836 families (Xue & Nation, 1984), Coxhead compiled a 3.5 million word corpus containing journal articles, textbooks and other academic writings from four disciplines: Arts, Commerce, Science, and Law. Using the Range software developed by

Heatley and Nation (1994), word families were selected for inclusion on the AWL if they 1) were not present on the GSL (specialization), 2) occurred in the corpus as a whole at least 100 times (overall frequency), and 3) occurred in each discipline or subcorpus at least 10 times (range). The result was a list of 570 word families covering approximately 10% of the academic corpus, now referred to as the AWL. The new list's coverage of a similarly sized corpus of literary fiction amounted to just 1.4%, indicating that the list does indeed contain words that are more prevalent in academic discourse. Coxhead argued that the manageable size of this list provided teachers with a useful guideline concerning the most useful target vocabulary to teach for improving the reading comprehension and written production of university students.

Building on the AWL, Coxhead and Hirsh (2007) focussed on an 875,000 word pilot science corpus to determine whether it was possible to create a Scientific Specific Word List (SSWL) that would provide additional coverage beyond the GSL and the AWL. Using similar overall frequency and range criteria as Coxhead (2000), they were able to identify a further 318 word families providing coverage of 3.89% in a 1.7 million word corpus of engineering, computer science, chemistry, biology, and physics texts. A similar procedure was replicated by Konstantakis (2007), who created a Business Word List (BWL) of 560 word families covering 2.79% of a 600,000 word corpus of business text books. It is interesting to note that neither the SSWL nor the BWL reached their intended coverage targets of 5%.

Although this approach seems to produce word lists that have reasonable text coverage, there are certain limitations to this method. Some have noted that not all university students have good knowledge of the first 2000 words on the GSL, implying that even if students learn vocabulary appearing on other specialized lists such as the AWL or BWL, they will continue to struggle for adequate comprehension of academic texts. Researchers wishing to address these students with more limited vocabulary knowledge therefore support the inclusion all word families meeting the range criteria, including those appearing in the top 2000 most frequent families. Ward (2009), for example, chose to create a Beginner Engineering List (BEL) from a 271,000 word corpus intended for Thai engineering students who had serious academic difficulties. He used texts from five engineering subdisciplines (e.g., chemical, mathematical, etc.) establishing a selection criterion of five occurrences per subdiscipline for a total minimum frequency of 25 occurrences. He found that his list of just 299 lemmas (210 families) provided coverage of between 17% and 21% of

a test engineering corpus while it provided only 5% coverage of a general language corpus. Of these lemmas, 132 appear in the GSL's 1000 most frequent families, accounting for the high coverage attained.

A second criticism is that the GSL itself, while still widely used because of its frequency information concerning the multiple meanings within each word family (Nation & Waring, 1997), is an outdated list omitting words such as *computer*, *online* or *job* (Eldridge, 2008) and may, therefore, not be a suitable representation of the words to be learned by the tech savvy students of today. A final limitation is that the nuanced context-specific meanings of certain terms are not differentiated in the specialized lists described above, giving rise to a potential source of confusion for learners (Hyland & Tse, 2007). For these reasons, some researchers have advocated a more holistic approach relying on the relative word frequency of lexis in a specialized corpus as compared to a more generalized corpus or word list in order to create their specialized lists (Coxhead & Hirsch, 2007).

Corpus Comparison

In this second approach to list creation, word types or families are included in the word list if they are significantly more frequent in a specialized corpus than in a corpus of more general texts or a list generated from a general corpus (Coxhead & Hirsh, 2007). In this approach, all specialized words, including those in the first 2000 most frequent families, are identified using electronic "term extractors" (Chung & Nation, 2004) that use statistical measures to calculate relative frequency. These measures take into account not only the frequency of words in the corpus, but also the specificity of their use, providing a more complex picture of specialized vocabulary profiles.

Chung & Nation (2004) evaluated this method using a 55,000 word anatomy book. They first reduced the lexis occurring in the corpus to a list of 876 lemmas, and submitted it to two health professionals, who rated the lemmas on a specialization scale. When the rating was complete, 226 lemmas were identified as technical vocabulary, or terms. The frequencies of the 876 lemmas as tallied in the text were then entered into a computer program and compared against the frequency profile of a 2 million word general language corpus. The authors found that the program was able to identify approximately 87% of 226 lemmas deemed to be technical lexis, indicating that this technique is highly efficient in producing accurate specialized word lists that reflect the intuitions of trained professionals.

Chujo and Utiyama (2006) used nine different statistical measures to compare a master word list for commerce and finance to a general word list in order to identify the top 500 most useful terms for learners studying commerce and finance. The commerce list of 2973 lemmas was compared to a list of the 13,994 most frequent lemmas as identified in the BNC. A different specificity score was attributed to each commerce term depending on the statistical procedure selected, creating nine different top 500 lists. The terms on each top 500 list were then analyzed to investigate where they fell in the standard BNC frequency bands (0-1000 most frequent words families, 1000-2000, etc.). The results showed that these lists differed significantly in the terms they included. *Raw frequency, cosine* and *complementary similarity measures* (CSM) produced a list of words remaining essentially within the first 2000 word families. *Chi square, Yates* and *loglikelihood* measures, on the other hand, produced lists that contained a greater proportion of vocabulary from 3-6000 frequency range, while *Mutual Information* and *McNemar* measures produced the largest number of infrequent words, with around 23% of lemmas landing in the 7000-13000 frequency bands. The authors concluded that statistical procedures for calculating term specificity for L2 teaching should be selected depending on the intended audience of the list, with *loglikelhood, chi square* and *Yates* measures being the most suitable for most intermediate level learners.

A limitation of both the layered and comparative approaches is that neither considers the important role played by multiword units, which may account for a significant portion of specialized texts (Biber, 2002; Jablonkai, 2010). Liu (2012) for example, provided a list of 226 frequent multiword units identified in a large-scale academic corpus, approximately two thirds of which had frequencies equal or higher to those included on the AWL. Examples include the multiword expressions *in fact*, *such as, according to, consist of, in addition,* and *participate in.* In their study of spoken and written academic discourse, Biber and Barbieri (2007) also found that multi-word units, referred to as lexical bundles, were highly prevalent, and especially frequent in institutional documentation such as course syllabi.

From this review it becomes clear that the method to be adopted when creating specialized lists depends on a number of key variables. Firstly, the researcher must consider the intended audience of the list, which will dictate not only the nature of texts included in the reference corpus, but also the choice of selection criteria in terms of range and overall frequency. Secondly, it is important to consider the overall aim of the list.

If it is to provide ample text coverage through the identification of the most frequent vocabulary, then selection based on frequency is appropriate. However, if the goal is to provide a list of vocabulary representative of a domain, field or text genre, specificity analyses employing corpus comparison techniques are likely more suitable. In addition, designers must take into account the importance of multiword units and whether they will be considered for inclusion. Finally, a decision must be made as to whether the list is based on word families or on lemmas. The family approach results in a shorter list but it gathers multiple meanings under a single form, raising the question as to whether a learner who knows the word *contract* is aware that one can contract a disease, sign a contract, work as a contractor, write contracted forms like *won't* and so on. The lemma approach treats these forms separately, but the downside is that a larger list will be necessary to provide ample text coverage.

PROJECT GOALS

For the project reported in this paper which examines the lexis of university websites intended for international students, the aim was first to discover whether or not we can reasonably expect English L2 university students to adequately comprehend these websites and if not, whether there is a list of vocabulary that would contribute significantly to helping them achieve adequate comprehension. The layered approach described in Coxhead and Hirsh (2007) and Konstantakis (2007) was adopted to answer these questions for two reasons. Firstly, students having passed the entrance examinations required for university in Canada can reasonably be expected to have knowledge of the most frequent 2000 word families of English. Secondly, the goal of the study is to potentially provide a word list which would complement students' previous knowledge rather than provide a complete lexical portrait of these websites. The research questions for this study are as follows:

1) Does knowledge of the first 2000 most frequent word families as identified in the first and second BNC frequency bands provide 95% coverage of a corpus of Canadian university websites and electronic literature targeted at international students?[1]

[1] The 95% coverage figure was selected in this case instead of the 98% target recommended by Schmitt, Jiang & Grabe (2011) because we believe that many obscure

2) If not, is there an identifiable specialized vocabulary associated with these texts which would allow them to achieve 95% coverage of the corpus?

METHOD

Corpus

The corpus of approximately 147,000 running words is composed of electronic texts that L2 English students would likely refer to on university websites. The following are some of the sources and topics included in the corpus:

- International student centre websites
- International student orientation handbooks and guides
- International student association events and news
- General admissions information and academic regulations
- Immigration, health, and tax information for international students
- English as a second language course descriptions
- Introductions to international student life including weather, housing, working, etc.

The texts were sampled from the websites of English Canadian Universities in four provinces. This choice was made in order to avoid biasing the data to one particular university's jargon and to be representative of Canadian universities in general. Table 1 shows the distribution of words in each university subcorpus.

Table 1. Corpus Composition

University	Province	Running Words
Concordia University	Quebec	37,450
Dalhousie University	Nova Scotia	35,800
University of Toronto	Ontario	36,350
University of British Columbia	British Columbia	37,550
Total		147,150

terms are explained within the texts themselves or at the very least links are provided to other websites with more detailed explanations. We recognize this is a limitation of our study.

To prepare the corpus for analysis, all postal and email addresses, websites, telephone numbers and lists of organizations or store names were removed in order to facilitate item recognition by the analysis software and to avoid artificially inflating the word count. Text from tables was only included if it contained information expressed in complete sentences, for example, when step-by-step instructions for an application procedure were provided. Lists of services, necessary forms, payment methods, etc. and headings were also judged to include relevant vocabulary for the L2 learner and were thus included in the corpus without bullets or numbering. Names of forms involving letters and numbers (e.g., k12.pdf) were removed. Finally, all incorrectly spelled words were corrected (e.g., activites) and the file was saved in plain text format.

Analysis

In order to answer the first research question, the corpus was submitted to the Web Vocabprofiler (Cobb, 2012a) on the Compleat Lexical Tutor website (see lextutor.ca). The software, based on the Heatley & Nation (1994) classic Range software, calculates the coverage of established word lists such as the GSL or AWL in a given corpus. In Cobb's version, it is also possible to examine a corpus against the top 20 frequency bands as identified in the BNC. Being that the BNC frequency band lists are more up to date than the GSL and the output provided by the BNC interface is more detailed, this option was selected.

The text was uploaded to the BNC profiler and submitted to a preliminary analysis. The list of offlist words (i.e., words not appearing in the 20 BNC lists) was examined to identify any typographical errors and to determine if words such as proper nouns should be recategorized as 1000-level words because of their low learning burdens. Subsequently, all acronyms (e.g., TOEFL, ITS, RCMP, SPCA), place names (e.g., Toronto, Bloor, Quebec), institution names (e.g., Concordia, UBC), people's names (e.g., Pam, George) as well as names of products, stores, or websites (e.g., Mac, Safeway, Facebook) were recategorized. In addition, words in other languages, such as French, were recategorized. A few exceptions of proper nouns and acronyms that were not recategorized include Skytrain and Metro, FYI, GPA, DVD and VCR. This decision was made because the acronyms and expressions were not made explicit in the text. The corpus was then inputted into the profiler a second time. Below in table 2 are the BNC coverage results.

Table 2. Coverage of the BNC Top 2000 Families

Frequency Band	Tokens	Types	Families	Coverage (% of tokens)
Top 1000 list	116,239	2,476	917	78.15
Top 2000 list	14,168	1,491	715	9.53
Recategorized list	4,291	625	-	2.92
Total coverage	134,698	4,592	1,632	90.60

As the table shows, the first 2000 most frequent words families combined with the recategorized acronyms and proper nouns cover approximately 90.6% of the corpus. To ensure this analysis was similar to the traditional GSL/AWL analysis, the corpus was also submitted to the classic Web Vocabprofile analysis. Results are presented in the table below. As can be seen by the total coverage figure of 90.64%, the results are remarkably similar.

Table 3. Coverage of the AWL and GSL

Lists	Tokens	Types	Families	Coverage (% of Tokens)
GSL 1-1000	107,986	2,173	880	73.37
GSL 1001-2000	9,328	960	519	6.34
AWL Words	11,783	1,045	446	8.01
Recategorized list	4291	625	-	2.92
Total coverage	133,388	4,803	1,845	90.64

In both cases, the target 95% text coverage is not reached, meaning that students with knowledge of only the 2000 most frequent word families in the BNC, or knowledge of the GSL and AWL, would not be able to attain adequate comprehension of these university websites.

Word list

Given the result reported in the section above, I attempted to identify a list of words that would cover the remaining 4.4%, a list that if learned by students would allow them to reach the threshold of 95%. In order to construct a list that would be representative of the corpus as a whole, three criteria based on those used in Coxhead (2000) and Coxhead and Hirsh (2007) were established for item selection. Firstly, the item could not be in the BNC top 2000 word families, it had to occur a minimum of seven times in the corpus as a whole and had to be present in at least three of the four subcorpora. While there is no established procedure for determining

the ideal number of occurrences necessary for list inclusion (Ward, 2009), it seemed reasonable to choose a number that was both adapted to the limited size of the corpus and that would produce a manageable list of candidates. After several trial tests, a cut off of seven occurrences proved to provide this optimal balance.

To extract items corresponding to these criteria, I used the Range analysis (Cobb, 2009), which compares lexis across corpora. Once the four subcorpora were submitted for analysis, Range automatically generated a list of 285 tokens matching the selection criteria. I then manually eliminated all place names (e.g., Montreal, Canada) and acronyms (e.g., ESL, TRV) as per previous word list research (Konstantakis, 2007; Coxhead, 2000). The list was subsequently collapsed into lemmas by grouping singular and plurals (graduate and graduates) as well as verb forms (e.g., submits, submitted, submitting) under a single headword. The choice was made to keep lemmatized types as list headwords rather than using word family headwords because, as suggested by Ward (2009) and Hyland and Tse (2007), when types were collapsed into families (i.e., groups with the same lexical stem such as academic, academy, academia), their specialized meanings were often masked. For example, the family headword for the extremely frequent token orientation is orient, which was not observed in the corpus. This procedure yielded a list of 226 headwords occurring approximately 6482 times throughout the corpus. This list will be referred to hereafter as the International Student Word List (ISWL).

Comparing the cumulative frequency figure for the ISWL (6482 occurrences) as supplied by the Range analysis to the total number of tokens in the corpus (147,150), we can estimate that items on the word list cover approximately 4.4% of the website corpus. Thus, with knowledge of the words on the ISWL and the top 2000 BNC word families, students would achieve the 95% coverage target for adequate comprehension of texts. The word list can be found presented in alphabetical order in Appendix A.

When words appearing on the AWL and the GSL were subtracted from ISWL (approximately one third of the list), the coverage figure drops to 3.34% of the corpus, a figure similar to that obtained by Coxhead and Hirsh (2007) in their analysis of a science specific corpus.

DISCUSSION

Results of the analyses described above show that the 2000 most frequent word families in the BNC cover only 90.6% of the student information corpus. The International Student Word List (ISWL), which consists of 226 lemmatized word types, covers an additional 4.4% of the corpus, allowing a total coverage of 95% to be reached.

While BNC coverage results alone did not reach the 95% target, the almost 91% rate is still good news for students who have a solid foundation in basic English. With 78% of word families falling within the first frequency band, students are likely to be familiar with a large number of words in this corpus. The nearly identical coverage figures obtained using the combined GSL and AWL confirm the assumption that the majority of this corpus's lexis is relatively simple.

To understand exactly to what extent knowledge of the ISWL would increase the comprehensibility, a portion of sample text was taken from the corpus and all words not appearing in the BNC top 2000 word families were removed. This was then compared to a text in which ISWL words were included. In figure 1, it is possible to see the results of this comparison.

This text shows the advantages afforded to learners who know words on the ISWL. Words such as campus and tuition are central to the message of these texts. They are also culturally loaded words, referring to specific North American academic practices. This means they probably present a high learning burden for students, who may be confused by partial meaning correspondences or may simply be unfamiliar with the concept entirely because it does not exist in their home culture. This might be the case for campus, for example, a concept which is completely lacking in some countries' university traditions.

The high density of the text in figure 1, as suggested by Biber (2002), may also present a challenge for students. When a multi-word analysis of the corpus was conducted using N-Gram (Cobb, 2012b), results showed that repeated three word strings accounted for nearly 27% of the corpus, while a comparative analysis of one million running words taken from the Brown corpus of general American English (Francis & Kucera, 1979) showed that three word strings accounted for only 17%. This figure points to a high proportion of repeated formulaic language which may have specialized meanings unknown to students. For example, collocations such as study permit, international student, off(-)campus, on(-)campus, and health insurance were extremely frequent but could not be included

on the list because one or more of their constituents appeared in the BNC top 2000 families. Some research has shown that while collocations may present significant challenges for learners in production, they are often semantically transparent and do not cause significant comprehension difficulties (Laufer & Girsai, 2008). Since students are not likely to write texts such as those in the corpus, the decision was made to view the ISWL as a list of words to be mastered for comprehension rather than production and thus collocations were not included.

Figure 1. Coverage Comparison for a Sample Corpus Text

BNC top 2000 + proper nouns	Combined coverage with ISWL
Financial Information You must arrange funding before leaving for Canada. When applying for the CAQ and the Canadian Study Permit, you will be required to present evidence of sufficient funds for _____and living expenses. Please see _____ and Fees. Canadian _____regulations allow International students to work on or off_____. Temporary employment, however, is not considered a sufficient source of funding, except for graduate students employed as research or teaching assistants. If you are _____ by your _____, _____ Canada will allow him/her to work on or off_____. Contact the International Students Office for information on obtaining off-_____ employment.	**Financial Information** You must arrange funding before leaving for Canada. When applying for the CAQ and the Canadian Study Permit, you will be required to present evidence of sufficient funds for tuition and living expenses. Please see Tuition and Fees. Canadian immigration regulations allow International students to work on or off campus. Temporary employment, however, is not considered a sufficient source of funding, except for graduate students employed as research or teaching assistants. If you are accompanied by your spouse, Immigration Canada will allow him/her to work on or off campus. Contact the International Students Office for information on obtaining off-campus employment.

Note. Text retrieved from:
http://www.concordia.ca/admissions/undergraduate/admission-requirements/international-requirements/

It is also interesting to observe the high proportion of proper nouns and acronyms. Nearly 3% of the text is composed of these words, 14 of which were so frequent, they conformed to the selection criteria for ISWL (e.g., CAD – Canadian dollars, HST- Harmonized Sales Tax). While these

acronyms are generally made explicit in the text, they may significantly increase the cognitive load for students if they are repeated without explanation.

For the purposes of list creation, it would seem that this corpus is ideally suited given the extensive coverage provided by the relatively limited number of word types on the ISWL. This high coverage rate improves on the results reported by Coxhead and Hirsh (2007) and Konstantakis (2007), who constructed significantly larger family lists (318 and 560 respectively), covering 3.9% and 2.8% of their target corpora. The ease with which relevant vocabulary could be identified can be attributed to the narrow scope of texts to which this list applies, namely university websites addressing international students. The texts included in the corpus address a number of closely related topics and were remarkably similar across universities. Table 4 shows the word families with a frequency of over 20 in each subcorpus (excluding acronyms, proper nouns and the BNC top 2000). All words except those marked with an asterisk were selected for the ISWL. The number of words recurring across the three corpora, such as campus, graduate, eligible, visa, and immigration is striking. The clear similarities suggest that although the corpus is relatively small by today's corpus study standards, it is likely representative of other similar Canadian university websites. However, because the coverage figures could not be calculated using a novel corpus of different university website texts, it is impossible to determine whether this word list is effective only for the corpus used in this study, or whether this world list would be relevant for other University sites.

Table 4 . Frequent Lemmas Across Subcorpora

U.of Toronto	#	UBC	#	Dalhousie	#	Concordia	#
campus	164	campus	196	academy	186	immigration	134
academy	163	graduate	111	faculty	136	campus	84
faculty	136	academy	92	graduate	119	academic	75
graduate	90	online	80	campus	113	visa	73
enrol	75	eligible	65	orient	64	deadline	53
eligible	65	faculty	57	undergrad	57	website	51
submit	63	undergrad	38	immigrate	50	graduate	49
immigrate	48	globe	37	online	40	undergrad	43
undergrad	44	submit	36	visa	35	tuition	41
visa	43	enrol	35	registrar	35	apartment	39
online	36	expire	29	eligible	34	proof	39
refund	36	passport	29	calendar	33	eligible	37

deadline	32	visa	27	bachelor	32	certificate	32
exempt	32	dental	25	transcript	30	passport	27
scholarship	28	professor	25	submit	30	opt	26
expire	26	clinic	24	dismiss*	29	workshops	25
mba*	25	transition	24	profession	27	exemption	23
proof	25	immigrate	22	architecture	24	online	23
approximate	21	harass*	22	cumulative	22	submit	22
origin	21	participate	22	buddy*	21	handbook	20
passport	21	counsel	21	dean	21	embassy	20
registrar	21					verify	20

Note. The asterisk marks words not appearing in the ISWL because of range criteria.

CONCLUSION

Overall, this project has uncovered that Canadian university websites, even those specifically addressing L2 speakers of English, are probably not entirely comprehensible to international students with knowledge of only the first 2000 most frequent words families as observed in the BNC. Building on the method used by Coxhead and Hirsh (2007), we have attempted to identify a complementary list of items that appear frequently in website corpus in order to attain a total combined coverage of 95%. The result is the ISWL, a list of 226 headwords covering approximately 4.4% of the corpus.

The ISWL, while probably not ideal for teaching in the ESL classroom, could potentially be employed as a guide for selecting words to be glossed in handbooks or on websites intended specifically for international students. Making definitions of these terms available to students through linking to online definitions or margin glosses could be of significant help to students who are new to the Canadian university system and unfamiliar with some of these culturally loaded concepts.

Limitations and Future Directions

While the results of this analysis are encouraging and have identified a relatively small number of very productive words, this methodology has its limitations. Firstly, the use of automatic analysis can sometimes categorize items incorrectly, skewing the coverage figures upon which this study has been based. A good example of this is the acronym SIN, categorized by the Vocabprofiler in the second BNC frequency band, but which actually refers to Social Insurance Number rather than a bad deed.

Secondly, this method assumes knowledge of the first 2000 BNC words families whereas it is highly plausible that students are either not familiar with certain word families or are unaware of the specialized uses made of them in the university context. In addition, many of the words included on ISWL are likely already known to students, particularly those related to the internet such as online, download, and website. Therefore users of this list must employ its contents critically, using their judgement when selecting items for teaching or glossing. Finally, the ISWL does not include multiword strings, which have been shown to be highly productive in this corpus.

In the future, ISWL coverage must be examined in a comparable corpus of university texts from Canada and other English speaking countries, in addition to a general language corpus in order to establish its level of specialization. The corpus should also be examined qualitatively to assess whether the words on the list are explained or made explicit for students within the texts themselves, which would substantially increase the likelihood of learner comprehension. Finally, a thorough study of the multiword strings in the corpus should be undertaken, with special attention paid to collocations or expressions that present a substantial learning burden specifically for comprehension. Hopefully, future studies will address these issues; we also hope that this initial study will be useful to researchers, educators, and website designers in the crucial endeavour of making the information more accessible to international students.

ACKNOWLEDGEMENTS

We would like to thank the two anonymous reviewers for their comments on earlier versions of this manuscript, as well as Tom Cobb for the use of the Lextutor website.

REFERENCES

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

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.

British National Corpus, (Version 3, BNC XML Edition) (2007). Oxford: Oxford University Computing Services. Available at http://www.natcorp.ox.ac.uk/

Chujo, K., & Utiyama, M. (2006). Selecting level-specific specialized vocabulary using statistical measures. *System, 34*, 255-269.

Chung, T. M, & Nation, I.S.P. (2004). Identifying technical vocabulary. *System, 32*, 251-263.

Cobb, T. (2009). Range (Version 2) [Web-based corpus analysis software adapted from Heatley & Nation, 1994]. Available at http://www.lextutor.ca/range/

Cobb, T. (2012a). WebVocabprofile (Version 3). [Web-based corpus analysis software adapted from Heatley & Nation, 1994]. Available at http://www.lextutor.ca/vp/eng/

Cobb, T. (2012b). N-Gram Phrase Extractor (Version 5.1) [Web-based corpus analysis software]. Available at http://lextutor.ca/tuples/eng/

Coxhead, A., & Hirsch, D. (2007). A pilot science-specific word list. *Revue française de linguistique appliquée*, *12*(2), 65-78.

Eldridge, J. (2008). "No, there isn't an 'academic vocabulary,' but…", *TESOL Quarterly, 42*(1), 109-113.

Francis, W., & Kucera, H. (1979). Brown Corpus of American English. Rhode Island: Brown University, Providence.

Heatley, A., & Nation, P. (1994). Range. Victoria, NZ: University of Wellington.

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly, 41*(2), 235-253.

Laufer, B., & Girsai, N. (2008). Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics, 29*, 694-716.

Liu, D. (2012). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes*, *31*(1), 25-35.

Jablonkai, R. (2010). English in the context of European integration: A corpus-driven analysis of lexical bundles in English EU documents. *English for Specific Purposes, 29*, 253- 267.

Konstantakis, N. (2007). Creating a business word list for teaching business English. *ELIA, 7*, 79-102.

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

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

Nation, I. S. P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt, & M. McCarthy, M. (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6-20). Cambridge: Cambridge University Press.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal. 95*(1), 26-43.

Ward, J. (2009). A basic engineering English word list for les proficient foundation engineering undergraduates. *English for Specific Purposes, 28*, 170-182.

West, M. (1953). *A General Service List of English Words*. London: Longman.

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

APPENDIX A

INTERNATIONAL STUDENT WORD LIST (ISWL)

abroad	currency	graduate	opting	scan
academic	curriculum	graduation	oral	scholar
accompanying	customs	grocery	orientation	scholarship
administer	deadline	guideline	outline	semester
administrative	dean	handbook	outstanding	seminar
airline	debit	hesitate	overview	shuttle
alternate	dedicated	host	participate	signature
alumni	dental	ID	participation	smooth
anxiety	departure	immigration	passport	solely
apartment	deposit	incoming	peer	spouse
approximately	designated	ineligible	personalized	stream
architecture	dial	informal	pharmacy	submission
athletic	dining	innovative	PhD	submit
atmosphere	diploma	inquire	photocopy	subsequent
authorization	diverse	integral	physics	summary
authorized	diversity	integration	plagiarism	supervisor
bachelor	download	integrity	port	supplemental
border	downtown	intensive	portal	surrounding
bulletin	duration	intercultural	practitioner	taxi
bursary	electronic	internet	premium	tenant
cable	eligibility	internship	prerequisite	terminal
calendar	eligible	lab	prescription	timetable
campus	email	laboratory	primary	transcript
certificate	embassy	laundry	prior	transit
certified	enhance	lease	probation	transition
checklist	enrol	listing	professor	translation
classmate	enrolment	lounge	proficiency	tuition
click	equivalent	mandatory	proof	tutor
climate	essay	media	province	tutorial
clinic	evaluation	mentor	provincial	tutoring
component	exceed	metro	publication	undergraduate
comprehensive	exempt	mild	receipt	unique
compulsory	exemption	ministry	recreation	upcoming
conduct	expire	multicultural	recreational	upper
confidential	expiry	multiple	refugee	urban
consecutive	explore	network	refund	verification
consent	fabric	nominated	refundable	verify
consist	faculty	nominee	registrar	via
consulate	fare	notification	reimburse	visa
convenience	fax	notify	reimbursement	vocabulary
convenient	federal	numerous	reputation	wealth
coordinator	format	occupation	restaurant	web
core	frequently	ongoing	revenue	website
counselling	global	online	review	wireless
courier	GPA	opt	roommate	workload
				workshop