The Lexical Breadth of Undergraduate Novice Level Writing Competency

Scott Roy Douglas
University of British Columbia, Okanagan Campus

Abstract

This study builds on previous work exploring reading and listening lexical thresholds (Nation, 2006; Laufer & Ravenhorst-Kalovski, 2010; Schmitt, Jiang, & Grabe, 2011) in order to investigate productive vocabulary targets that mark successful entry-level undergraduate writing. Papers that passed the Effective Writing Test (EWT) were chosen to create a corpus of novice university level writing (N = 120). Vocabulary profiles were generated, with results indicating the General Service List (GSL) and the Academic Word List (AWL) cover an average of 94% of a typical paper. Further analysis pointed to 3,000 word families and 5,000 word families covering 95% and 98% respectively of each paper. Low frequency lexical choices from beyond the 8,000 word family boundary accounted for only 0.6% coverage. These results support the frequency principle of vocabulary learning (Coxhead, 2006), and provide lexical targets for English for Academic Purposes (EAP) curriculum development and materials design.

Résumé

Cette étude s'appuie sur des travaux antérieurs qui explorent les niveaux lexicaux pour la lecture et l'écoute (Laufer et Ravenhorst-Kalovski, 2010; Nation, 2006; Schmitt, Jiang et Grabe, 2011). Elle a pour but d'étudier les niveaux de production lexicale qui marquent l'écriture à l'entrée à l'université anglophone. Pour créer un corpus d'écriture de niveau universitaire novice, 120 articles qui ont passé le Effective Writing Test (EWT) ont été choisis. Des profils de vocabulaire ont été générés et les résultats signalent que la General Service List (GSL) et la Academic Word List (AWL) couvrent une moyenne de 94% d'un document typique. En plus, 3 000 familles de mots et 5 000 familles de mots couvrent 95% et 98% respectivement de chaque article. Les choix de basses fréquences lexicales au-delà de la limite de 8 000 mots ne représentaient que 0,6% de la couverture. Ces résultats appuient le principe fréquence de l'apprentissage du vocabulaire (Coxhead, 2006) et fournissent des niveaux lexicaux pour les programmes d'anglais à des fins académiques.

The Lexical Breadth of Undergraduate Novice Level Writing Competency

Introduction

Lamentations about the poor writing skills of university students are plentiful in the popular press (for examples see Gilmour, 2010; Kelley, 2010), with particular accusations aimed at the "meagre" vocabularies of undergraduate students (Wilce, 2006). In Canada, the focus on supposedly poor writing skills has implications for the increasing numbers of linguistically diverse students enrolling in English medium institutions of higher education. The demographic make up of Canadian universities is changing. This demographic shift is due, in part, to high immigration levels that have led to a change in the languages people speak on a daily basis. For example, in cities such as Vancouver, 31% of the population reports speaking an immigrant language at home, with Punjabi, Chinese, and Tagalog being the most common (Statistics Canada, 2012). As the children of these families make their way through the K-12 school system, their families have strong aspirations for them to continue on to post-secondary education (Grayson, 2008; Krahn & Taylor, 2005). This is evidenced by the relatively high percentage of high school graduates from linguistically diverse backgrounds in the first year admissions of universities such as the University of Calgary (Roessingh & Douglas, 2012).

This increase in domestic students from non-English-speaking backgrounds is coupled with greater numbers of international students coming to study in Canada. For example, the British Columbia (BC) government has set ambitious goals to increase the number of international students coming to BC by 50% in just four years (Ministry of Advanced Education, Innovation, and Technology, 2012). In fact, the number of international students in Canada has increased over 60% since 2004, with Canada as a whole welcoming a record number of 100,000 international students in 2012 (Citizenship and Immigration Canada, 2013). As university campuses welcome greater numbers of both international students and newcomers with developing levels of English language proficiency, improved understanding of the actual lexical demands of entry-level university writing competence is vital for appropriate vocabulary-focused writing instruction and support.

What university level writing competence entails is shrouded in misconceptions around the developmental process of learning how to write in ever increasingly decontextualized and cognitively challenging academic tasks. These tasks include multi-page research papers incorporating references from peer reviewed journals, essay exams, and large-scale standardized writing assessments. Even undergraduate students from English-speaking backgrounds do not arrive in first year university studies as the ready-made users of academic English writing skills hoped for by their instructors. All undergraduate students are still learning how to write for academic purposes. Thus, the standards of accomplished writers in the academy, such as the authors of the texts used in undergraduate courses, are unfairly applied at entrance to new students enrolled in a program of post-secondary studies. Rather, four more years of development lie in front of novice undergraduate writers as they refine their writing skills and hone their ability to deploy the vocabulary needed to convey to their instructors the written evidence of their learning.

This is even truer for multilingual and developing users of English who have been deemed admissible to university studies. Applying standards far beyond what students can

be expected to produce is neither encouraging nor productive. In response to this danger of mismatched standards, the current study is a quantitative corpus-based inquiry that investigates vocabulary usage in writing samples produced by novice academic writers on entrance to university in order to uncover realistic lexical goals for developing users of English. The research problem at hand concerns the vocabulary elements of sound curriculum design for developing academic writers from non-English-speaking backgrounds. Curriculum design starts with first defining student goals before establishing assessments that will provide evidence of students reaching those goals and before designing the learning experiences that will foster the skills students need in order to perform well and undergo the assessments (Wiggins & McTighe, 2005). However, the productive vocabulary of novice undergraduate writers remains to be quantifiably verified. Thus, there is a danger that assessments and learning experiences are being designed without a clear understanding of a crucial piece of curriculum design: knowing the lexical goals.

The current study is meaningful in that its exploration of the lexical breadth of novice university level writing competence contributes to a better understanding of the lexical goals that English language development programs can set in order to prepare students for English-medium higher education. By providing obtainable goals and clarifying the lexical developmental process at work in undergraduate students as they begin to learn to write for their academic programs of study, the results could potentially help multilingual and developing users of English have a better chance at building up the productive vocabulary that accompanies academic success.

Relevant Literature

Academic Language Proficiency

The language proficiency that accompanies success in academic studies has been characterized by Cummins (1981) as language that is used in increasingly context reduced and cognitively challenging school settings. This is a situation in which students become more and more dependent on their knowledge of and ability to use language in order to access ideas and express meaning. Both Cummins (1981) and Roessingh (2006) likened academic language proficiency to the hidden mass of an iceberg. As opposed to the conversational language proficiency employed in day-to-day communicative interactions, much of the language that is needed for school success lies below the surface, providing the foundation necessary to engage in academic endeavours.

Underlying Proficiency

In his Common Underlying Proficiency (CUP) model, Cummins (1981) saw that for developing English language users, the context reduced and cognitively challenging aspects of both their first language (L1) and their additional language (L2) (i.e., academic language proficiency) can be seen as interdependent across both languages. Both the L1 and the L2 promote the proficiency underlying the two languages. This is typified by Cummins' model of two overlapping icebergs. The proficiency shared by both languages lies below the surface features of each language providing for the shared academic language proficiency needed for cognitively demanding tasks. Figure 1 illustrates the CUP model.

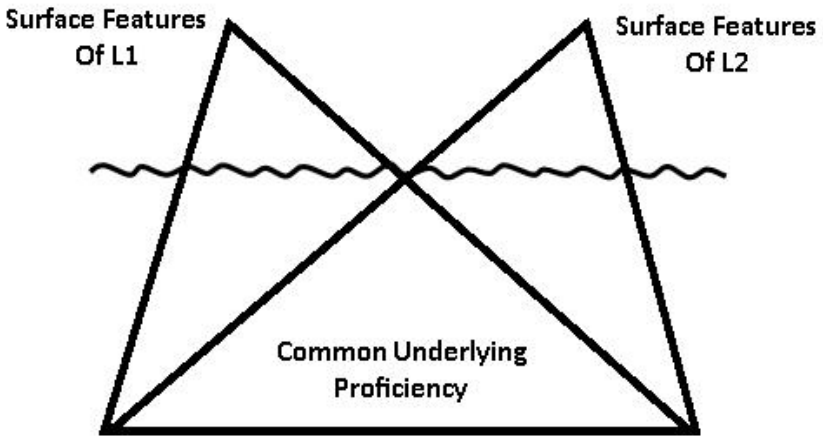

Figure 1. Common Underlying Proficiency Model of Bilingual Proficiency (Cummins, 1981, p. 24).

Vocabulary as an Underlying Variable of Academic Writing Competence

Roessingh (2006) has identified vocabulary as the underlying variable of academic language proficiency. Vocabulary is a key element for obtaining higher levels of academic language proficiency and the concomitant gains in academic outcomes. This is because there is an intimate relationship between vocabulary and academic writing. The vocabulary students employ enables them to express their learning and understanding in academic tasks that are progressively decontextualized and cognitively challenging. In fact, it is through the lens of their written work that students are typically evaluated in academic settings. As Nation (2008) has maintained, a skilled use of vocabulary demonstrates knowledge of a subject being studied, and the ability to use vocabulary effectively allows for the production of the writing evidence required for academic success.

Vocabulary use has a major impact on academic writing quality by contributing to generating, developing, and presenting ideas in meaningful texts (Engber, 1995; Grabe, 1985; McNamara, Crossley, & McCarthy, 2010; Raimes, 1983, 1985). Students lacking the range of vocabulary necessary to explore ideas have trouble making their thoughts concrete (Spack, 1984). However, as the ability to use vocabulary develops, this is very often accompanied by improved writing skills (Smith, 2003), with the size and range of a writer's available lexicon connected to the quality of a piece of writing (Brynildssen, 2000). Students with more vocabulary at their disposal are able to achieve higher evaluations of their writing (Nation, 2001) because a rich vocabulary has a positive effect on a reader (Laufer, 1994). Furthermore, lexical richness and higher ratings of essays correlate (Laufer & Nation, 1995), with simple lexical constructions having a negative impact on the quality of academic essays (Cobb, 2003; Hinkel, 2003). Highly rated vocabulary pulls up the general evaluation of a piece of academic writing, while poorly rated vocabulary pulls down the general evaluation of a piece of academic writing (Roessingh, 2008). Thus, an effective vocabulary contributes to producing the writing evidence that accompanies higher academic outcomes.

Vocabulary Breadth of Knowledge

Given the central importance of vocabulary, numerous studies have been carried out to quantify the amount of vocabulary known by adult native speakers. A common unit of measurement for counting vocabulary is the world family, with a word family being defined as including a base word plus all of its inflected and derived forms.

Zechmeister, D'Anna, Hall, Paus, and Smith (1993) pointed to converging evidence that undergraduate students know up to 17,000 word families. Zechmeister, Chronis, Cull, D'Anna, and Healy (1995) further maintained that older adults know about 22,000 word families, 5,000 more than typical college students. Nation (2001) contributed to these figures by asserting "educated adult native speakers of English know around 20,000 word families" (p. 9). By comparing educated adults and post-secondary level students, it seems that vocabulary acquisition continues throughout the higher education experience. This lexical growth establishes that the entry-level lexical targets for post-secondary studies are not necessarily the same as the exit-level vocabulary goals.

Going back through the childhood years, both Nation and Waring (1997) and Hart and Risley (2003) estimated that 5-year-old children know approximately 5,000 word families and add 1,000 word families to their lexicon for every year of education. Given Nation's (2001) "rough rule of thumb" (p. 9) that native speakers acquire about 1,000 word families each year, 13 years of school (including Kindergarten) would add 13,000 word families to a 5-year-old's initial 5,000 word families. From this, it can be estimated that an 18-year-old could be expected to know about 18,000 word families on leaving Grade 12, with more word families to be added as they engage in the content of their academic disciplines in post-secondary programs of study. This falls within estimates of university students knowing between 15,000 and 18,000 world families (Schmitt, 2010).

If the assumption is made that university-bound high school graduates know an estimated 18,000 word families, it is important to consider what it means to actually know a word as not all 18,000 of those word families are in regular use. Word knowledge is commonly seen on a continuum with passive or receptive vocabulary knowledge at one end of the spectrum and active or productive vocabulary knowledge at the other. Generally, passive or receptive vocabulary includes those words that can be remembered and understood when heard or read (Nation, 2001). Passive vocabulary is a set of vocabulary that includes the active vocabulary as well as those words which are not used actively because they may be only partially understood, they may be rarely encountered, or they may be avoided in use (Corson, 1995). Rising out of passive or receptive vocabulary is a generally smaller subset of active or productive vocabulary that can be used appropriately and accurately when spoken or written (Nation, 2001). Thus, for university-bound high school graduates, it is likely that their productive vocabulary output is going to be smaller than the estimated 18,000 word families that they possess.

Creating a Corpus

When investigating the breadth and depth of productive vocabulary knowledge, compiling corpora of writing samples provides useful data for analysis. The quantitative analysis of a corpus allows for research into language use that cannot usually be carried out by other methods (Biber & Conrad, 2001). Conrad (2005) saw the creation of a corpus as involving the grouping and electronic storing of authentic texts. McEnery, Xiao, and Tono

(2006) expanded on the definition by referring to how the sampled texts in a corpus should be representative of a language or language variety. Additionally, a corpus is typically made up of texts that are authentically written by users of language for real purposes (Conrad, 2005). The number of words needed for an effective corpus varies depending on the purpose of the corpus, with very large corpora not always being necessary as the best corpus size is based on the research questions being asked and the practical concerns of compiling authentic and representative language samples (McEnery et al., 2006).

Lexical Frequency Profiling

Of particular value in measuring a writing sample's productive vocabulary is an extrinsic measure of vocabulary size such as lexical frequency profiling. Lexical frequency profiling measures vocabulary distribution through the use of external frequency lists based on how frequently words are used in a language as represented by a large-scale corpus or collection of corpora. Laufer and Nation (1995) developed the Lexical Frequency Profile (LFP) to measure the percentage coverage of a text by different frequency levels of vocabulary. In the development of the LFP, the measurement was found to be stable across writing samples and able to differentiate between samples written with differing language ability. The LFP measure further correlated with an alternative vocabulary measure.

An online version of the LFP, Web VP, is based on Nation's VocabProfile and RANGE programs (Cobb, 2012; Heatley & Nation, 1994). Web VP quantifies the words used in a text using the General Service List (GSL) (West, 1953) and the Academic Word List (AWL) (Coxhead, 2000). By doing so, the coverage of a given text by the GSL (the 2000 most frequent words) and the AWL (a list of 570 academic words) is reported along with the percentage of the text not covered by either of those lists.

Along with Web VP, Cobb (2012) has developed the Web VP BNC-20 vocabulary profiler. The BNC-20 vocabulary profiler is based on 20 bands of 1,000 words each, with each band representing decreasing levels of lexical frequency in the British National Corpus (BNC). The percentage of coverage by each band is reported in the output of the BNC-20 profiler. The BNC-20 profiler can be used to demonstrate how far the vocabulary in a sample of text taps into less frequently occurring vocabulary bands, thus giving an indication of the productive lexical breadth found in a text.

Lexical Thresholds

Lexical frequency profiling can provide important information regarding the lexical thresholds that facilitate carrying out linguistic tasks, such as reading and listening. Nation (2006) came to the conclusion that independent reading and listening comprehension can take place at a 98% threshold. He estimated that at 98%, general readers would need to know 8,000 to 9,000 word families and listeners would need to know 6,000 to 7,000 word families. Turning to academic language, using the reading texts from a university entrance exam to explore the threshold requirements of academic texts, Laufer and Ravenhorst-Kalovski (2010) also explored the number of word families readers would need to know automatically in order to reach the text coverage thresholds necessary for adequate reading comprehension. Laufer and Ravenhorst-Kalovski's work indicated vocabulary goals for developing users of English aiming to read academic texts independently. They concluded that in order to read at an independent level, the lexical threshold is approximately 8,000

word families representing 98% text coverage. For students reading with guidance, the threshold is lower, with approximately 4,000 to 5,000 word families providing 95% text coverage. These thresholds are confirmed by Schmitt, Jiang, and Grabe (2011) who also postulated that readers should know about 98% of a text's vocabulary (8,000 to 9,000 word families) in order to read independently, while a 95% target would work for teacher supported texts.

Research Questions

While extensive work has been done exploring the lexical thresholds necessary to facilitate independent reading (Hu & Nation, 2000; Laufer, 1989; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006; Schmitt et al., 2011), little work has been done establishing the vocabulary thresholds that facilitate writing at a novice undergraduate level. However, investigating these productive thresholds is vitally important for setting lexical targets that realistically prepare developing users of English for academic success. Receptive vocabulary goals of 8,000 word families have been offered for independently coping with reading in English, but the question remains as to how many of those 8,000 word families should be actively known for productive purposes such as meeting the writing demands of first year undergraduate studies in English. This study sets out to investigate the vocabulary profiles that characterize entry-level academic writing competence by generating LFPs for novice university level writing competence as measured by the Effective Writing Test (EWT). This is done in order to explore the possibility of establishing lexical thresholds for novice university level writing competence. As such, the overarching research question for the study is: In terms of lexical breadth of knowledge, what characterizes the vocabulary of novice university level writing competence?

In answering the overarching research question characterizing the vocabulary of novice university level writing competence, a series of refining questions guided the research method:

1. What percentage of novice university level writing is covered by the 2,000 most frequent words in English (GSL), the AWL, and words not found on either of those lists?
2. How many of the BNC 1,000 word family frequency bands do students writing with novice university level writing competence access in order to cover 95%, 98%, 99%, and 100% of their total lexical output?
3. How many low frequency word families (higher than the 8,000 word family level) are found in novice university level writing?

Methodology

Participants

The location for this study was a large research-intensive university in Western Canada with a total enrolment of over 29,000 students. The study focused on first year students who sat the EWT between 2003 and 2006. The EWT is designed to measure university level writing competence. Students who pass the test with scores of *Marginally Satisfactory* or *Satisfactory* are deemed to have a level of writing competence sufficient to engage in the academic demands of undergraduate studies.

Students who took the EWT were chosen for inclusion in this study due to the ability to access the archives of the EWT. The EWT archives also provided a ready pool of independently assessed writing samples. As part of the EWT, students were asked to indicate whether they allowed their test papers to be used for anonymous research and educational purposes. Only papers for which students gave permission for use for research and educational purposes were considered for inclusion in the study. The corpus was built by sampling from those papers using purposive sampling techniques (Gay, Mills, & Airasian, 2012). A total of 120 student papers were pulled from the archives of the EWT. In order to be included in the corpus of novice university level writing competence, papers had to have received a passing score on the EWT (*Marginally Satisfactory* or *Satisfactory*). Purposive sampling created a corpus of papers the researcher believed to be representative of novice university level writing competence as measured by the EWT. Thus, a pool of 120 papers deemed to fulfill the requirements of novice university level writing competence were compiled to create a corpus of novice university level writing. Of the 120 papers, 64 were written by male students while 56 of the papers were written by female students. Students who reported English as their L1 wrote 117 of the papers, and for three of the papers, students did not provide their L1 information. The writers of the sampled papers came from the full range of academic disciplines, as seen in Table 1.

Table 1

Participant Programs of Study

Faculty	Participants
Communication and Culture	29
Science	19
Social Sciences	19
Engineering	16
Humanities	9
Fine Arts	7
Business	5
Nursing	5
Computer Science	4
Kinesiology	3
Education	2
Unclassified	2
Total Participants	120

The Effective Writing Test

The goal of the EWT is to assess a student's ability to demonstrate university entrance-level writing competence. The EWT requires that students write an expository essay of approximately 400 words during a 2 hr 30 min period. During the test, students may use a paper-based unilingual English dictionary. Writing topics are of a general academic nature, and students are given a choice of four writing topics to choose from. Questions are designed so that no specialized knowledge beyond that held by an average

high school graduate is required to answer the writing prompts. For example, one of the topics from the corpus is as follows: "Given all the evidence that cigarette smoking is harmful, why do people continue to smoke or take up smoking?" See Table 2 for an overview of general topics found in the current corpus.

Table 2

Range of Topics Included in the Corpus

Topic	Papers
Animal Welfare	8
Education	48
Entertainment	6
Health and Wellness	5
Politics	25
Sports	13
Technology	9
Work and Business	6
Total	120

In order to achieve reliable results on the EWT, each paper is blind scored independently by two assessors. If there is a disagreement in the score that would result in a paper failing, a third marker scores the paper. All assessors follow the guidelines in the *Assessors' Guide for the Effective Writing Test* (Effective Writing, 2003). As per these guidelines, each paper is marked with a detailed marking code (Effective Writing, 1993) that tracks performance in content, structure, paragraphs, sentences, grammar, word use, and spelling and punctuation. Error levels are set for each category to determine whether or not a student fails or half-fails a category. Based on the number of categories a student fails, papers are judged as *Unsatisfactory*, *Marginally Unsatisfactory*, *Marginally Satisfactory*, and *Satisfactory*. In order to be judged as *Marginally Satisfactory* or *Satisfactory*, a paper should not fail in more than one and a half categories. Students have to receive a score of at least *Marginally Satisfactory* in order to be judged as having entry-level university writing competence. It has previously been found in a study that compared EWT results and university transcripts (Douglas, 2010) that a similar population of students whose EWTs passed and were rated by one or more assessors as being *Satisfactory* went on to have higher grade point averages, fewer credits attempted and not earned, shorter lengths of program, and fewer incidences of being required to withdraw or being placed on academic probation compared to students whose essays did not pass the EWT. As a result, the main criterion for inclusion in the current study and judging whether a paper represented novice university level writing competence was that one or both of the assessors for each paper rated the paper as *Satisfactory*. Of the 120 papers included in the study, 25 papers received two scores of *Satisfactory*, 93 papers received one score of *Satisfactory* and one score of *Marginally Satisfactory*, and two papers received one score of *Satisfactory* and one score of *Marginally Unsatisfactory*. There was thus 21% agreement on the *Satisfactory* scores, and an overall 98% agreement on whether the papers chosen for inclusion in the corpus demonstrated entrance-level university writing competence.

Building the Corpus

The papers in this study were handwritten. As a result, each paper was digitized for inclusion in the corpus by typing the essays into Microsoft Word. At this stage of the digitization process, papers were also corrected for spelling errors and then saved in ASCII Text format so that the data could be read by the corpus analysis tools. The digitization of 120 papers resulted in a corpus of 62,309 running words. Each paper was an average of 519 words in length (M = 519, SD = 121.10). Because students could choose from four different topics of a general academic nature, papers were written on a variety of topics. Table 2 tabulates the range of general topics included in the corpus.

In order to determine if the differing topics were equally valid for inclusion in the corpus, the average percentage coverage of the GSL for each paper was determined. A simple analysis of variance (ANOVA) was calculated in IBN SPSS 20 to evaluate the relationship between topic choice and text coverage by the GSL. The analysis demonstrated that a topic's effect on text coverage by the GSL was not statistically significant, $F(7, 112) = 2.11$, $p = .05$. Thus, it was assumed that all of the topics included in the corpus equally contributed to an understanding of novice university level writing competence as any differences in the coverage of the GSL in papers written on different topics were not statistically significant.

Procedure

The first phase of the research method involved generating a vocabulary profile for each paper using online vocabulary profiling tools freely available on the Compleat Lexical Tutor website (Cobb, 2012). Similar to other lexical frequency profiling studies (e.g., Morris & Cobb, 2004) for each vocabulary profile generated, all proper nouns were categorized as belonging to the first 1,000 most frequent word families so that they would not appear as Off List vocabulary items or in the low frequency word family bands. In order to investigate the first refined research question regarding the GSL (K1-K2), AWL, and words not found on either of those lists (Off List), vocabulary profiles were generated for each of the sample papers in the corpus using Web VP v3 (Cobb, 2012).

The next refined research question focused on the ability to tap into the 1,000 word family frequency bands (K1, K2, K3, etc.) of the BNC in order to cover 95%, 98%, 99%, and 100% of each paper. Web VP/BNC-20 v3.2 (Cobb, 2012) generated the necessary BNC-20 vocabulary profiles.

Utilizing the same vocabulary profiles generated by the Web VP/BNC-20 v3.2, the number of low frequency word families in each sample paper were identified in order to answer the third refined research question investigating the number of word families beyond the K8 level found in the average sample paper. For each paper, the number of word families found in the 8,000 word frequency band (K8) and beyond (K8+) was counted and a ratio was calculated based on the number of K8+ word families divided by the total number of words and multiplied by 100 [(K8+/Tokens)*100]. The resulting ratio (K8+R) provided the number of K8+ word families per 100 running words of text in each writing sample.

Results and Discussion

Coverage by the GSL and the AWL

The writing samples of first year undergraduate students assessed to have novice university level writing competence revealed different levels of coverage by the GSL and the AWL compared to the academic texts students encounter at university. In the current study's corpus, the GSL covered, on average, 87.65% of each text, while the AWL covered, on average, 6.74% of each text. The remaining 5.61% of each text was covered by word families not found on either the GSL or the AWL. Results are reported in Table 3.

Table 3

Percentage Coverage GSL (K1-K2), AWL, and Off List (N =120)

	Minimum	Maximum	*M*	*SD*
K1-K2	77.62	93.97	87.65	3.16
AWL	1.17	18.21	6.74	2.60
Off List	0.60	11.80	5.61	2.24

The current study builds on the work of Coxhead's development of the AWL (Coxhead, 2000, 2011) in that it extends the investigation of the coverage of the AWL to novice university level writing competence. Coxhead's (2000) corpus of 414 academic texts including academic journal articles, book chapters, laboratory manuals, and university textbooks was put together to represent the reading materials that first year university students would likely encounter. In Coxhead's corpus, the GSL accounted for 76.1% of the academic corpus, and the AWL accounted for 10% of the academic corpus for a total of 86.1%. For papers in the current study, the GSL and the AWL accounted for a total of 94.39%. Thus, the academic texts students may eventually have to read as part of their studies contain higher levels of low frequency word choices, and the novice academic texts students produce contain higher levels of high frequency word choices. Student-generated texts rely more on the 2,000 most frequent words of English and employ less of a general academic vocabulary, as represented by the AWL, than the texts they will eventually read at university. In other words, students need a greater breadth of passive vocabulary knowledge for the reading tasks they encounter in their academic studies compared the breadth of active vocabulary knowledge needed to produce the initial academic writing tasks they face. As a result, productive vocabulary targets can be set at lower levels than passive vocabulary thresholds for reading. Of note to developing users of English who are bound for higher education in English, a solid knowledge of the first 2,000 most frequent word families of English (the GSL) along with the AWL (570 word families) accounts for just over 94% of the vocabulary they need for productive writing tasks of a general academic nature at the start of their undergraduate studies. A first step toward reaching the level of novice university level writing competence is having a full and automatic command of these 2,570 word families. This is a clear goal in reach of many linguistically diverse students.

Lexical Stretch

Results for the investigation into the number of word families required to reach threshold percentages of text in novice university level writing are reported in Table 4.

Table 4

BNC-20 Word Family Frequency Bands Required to Reach Coverage Percentages (N = 120)

	Minimum	Maximum	*M*	*SD*
95% Coverage	2	6	3.26	0.97
98% Coverage	2	12	5.31	1.47
99% Coverage	2	15	6.73	2.00
100% Coverage	6	20	11.69	3.49

On average, writing samples in the corpus contained word families that stretched just past the 11K mark. These data point to novice university level writing competence being marked by an overall ability to tap into between 11,000 and 12,000 word families. Thus, students in the study have a capacity to stretch their vocabulary production toward the 12,000 word family band. While not all of the word families in each frequency band of the BNC may be part of a novice writer's productive vocabulary, it seems that the reach into the lower frequency bands is a marker of novice university level writing competence.

In terms of the productive vocabulary stretch at entrance to higher studies, a more detailed analysis of the coverage provided by the BNC-20 frequency bands in the corpus reveals that the last 5% of an essay places the greatest lexical demands on the students' vocabulary. While the BNC-20 analysis shows that the ability to utilize lexical choices from the 3,000 to 4,000 word family bands is needed to cover the first 95% of the average writing sample, over 8,000 more word families are required to cover last 5% of the average writing sample. Figure 2 demonstrates this increase in lexical demand.

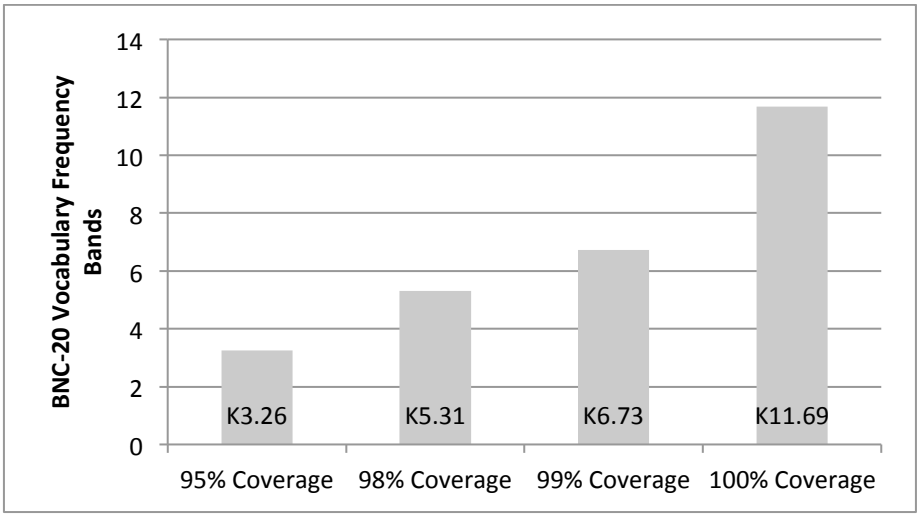

Figure 2. BNC vocabulary frequency bands and percentage coverage of text.

The last five percent of the average essay requires the most lexical resources from the novice academic writer. However, a closer look uncovers goals that can be established for developing users of English in preparation for academic studies. A goal of being able to productively use about 3,000 world families will adequately cover approximately 95% of the lexical output found in novice university level writing competence. From 95% coverage to 98% coverage is a difference of around 2,000 word families. Thus, developing users of English can begin to approach the vocabulary output found in novice university level writing competence by adding access to a further two frequency bands for an active vocabulary set of around 5,000 word families to cover about 98% of a written text at this level.

As a first step, a goal of achieving automatic productive capacity in the most frequent 3,000 world families of English will afford developing writers the ability to produce 95% of the lexical output of their peers writing with novice university level competence. By being able to produce 95% of the lexical output of their novice university level competent peers, developing writers will have the capacity to produce 19 out of 20 word families to which their more proficient peers have access. This ability leaves a lexical gap of 1 in 20 words of running text for which developing writers do not have an automatically available word choice in English. However, by achieving a goal of automaticity in the most frequent 5,000 word families in English, developing writers potentially have the ability to produce 98% of the lexical output of their university level competent peers. On reaching 98% of the lexical output that marks university level writing competence, developing writers now have access to 49 out of 50 running words of text found at this level. Thus, the lexical gap is only 1 in 50 word families.

Coming to university studies with 5,000 active word families would have the downstream effect of creating a framework for linguistically diverse students to begin meeting the writing challenges of novice university level studies. Developing users of English would only be missing about two possible word choices from every 100 running words of text, a manageable amount that can come from a linguistically diverse student's underlying language proficiency via translation without placing too much cognitive demand on the student. In other words, by automatizing 98% of the vocabulary needed to produce a satisfactory text, students can have recourse to their L1 for the other 2% of the essay's lexical output without over-taxing the lexical resources of the writer. The availability of a solid base of 5,000 word families will also provide quick access to words so that more cognition time is available for focusing on ideas as opposed to searching for vocabulary (Coxhead & Nation, 2001). Automatizing 5,000 word families is thus proposed as an entry-level threshold for the independent production of novice academic writing. As developing users of English engage in their academic programs of study, further vocabulary growth will be needed in order for students to keep up with the vocabulary growth of their undergraduate peers. However, setting obtainable goals will lead to the creation of a solid foundation on which to map new vocabulary encounters. As a result, by adding to the understanding of the receptive vocabulary thresholds of reading and listening (Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006; Schmitt et al., 2011), these productive vocabulary thresholds point the way to understanding the productive lexical requirements that facilitate novice university level writing.

Low Frequency Vocabulary

When looking at the usage of low frequency vocabulary in the average paper included in the corpus, low frequency word families, as defined as being from the K8 band and beyond, appear to be used sparingly. Table 5 shows the average number of lexical items from beyond the K8 band in 100 running words of text in novice university level writing.

Table 5

K8+ Word Families Per 100 Running Words of Text (N = 120)

	Minimum	Maximum	*M*	*SD*
K8+R	0.00	2.19	0.60	0.42

With a K8+ Ratio (K8+R) of 0.60 and the average paper having 519 running words, each paper would, on average, contain about three low frequency word choices. This can also be interpreted as one low frequency word choice for every 167 running words of text. With only one word in 167, or less than 1%, of an essay requiring a low frequency word choice, the relative rarity of low frequency vocabulary in novice university level writing again points to the possibility of multilingual developing users of English being able to tap into their underlying proficiency in their L1s to find and translate the low frequency vocabulary they need. By mediating their underlying proficiency, students from non-English-speaking backgrounds who have automatized a certain level of vocabulary will infrequently be stalled searching for a low frequency word choice.

Conclusion

Gaining university level writing competence is a developmental process. Most students do not arrive at university with a set of productive vocabulary equal to the lexicon used by scholars to produce highly developed levels of academic writing, such as the texts undergraduates likely encounter and have to read at university. Thus, understanding the threshold number of word families students can be expected to understand on entrance to higher education, as well as the threshold number of word families they can reasonably be expected to have easy access to in their writing, has important implications for multilingual and developing users of English.

By knowing the lexical thresholds that mark novice university level writing competence, vocabulary goals can be incorporated into the curriculum of both English for Academic Purposes (EAP) programs and developmental writing courses. Word family thresholds also inform writing assessment on entry to undergraduate studies. While certain levels of vocabulary are necessary to complete academic writing assignments successfully, reaching the total overall vocabulary output levels of more proficient writers is not a realistic goal. Instead, reasonable thresholds provide practical guidelines for EAP curriculum development.

The papers included in this study's corpus of novice university level writing competence point to word choices from approximately the first three 1,000 word frequency bands of the BNC as covering 95% of the lexis in an average participant's paper, and word

choices from approximately the first five 1,000 word frequency bands of the BNC as covering 98% of the lexis of an average participant's paper. The results further point to the relative rarity of word choices from beyond the 8,000 word frequency band of the BNC as these low frequency lexical items cover a mere 0.6% of an average text in the corpus of novice university level writing competence.

These findings support the utility of the frequency principle (Coxhead, 2006) for learning vocabulary. Coxhead has advocated concentrating on high frequency vocabulary items first because they are a fundamental part of language. Once effort has been invested in acquiring the most frequent word choices, then attention can be turned to lower frequency vocabulary items. Based on the findings in this study, vocabulary goals for productive use can be set for students studying English in order to gain entry to undergraduate studies at an English medium institution of higher education. These productive thresholds can be used as a guide when deciding on vocabulary targets in a program of studies for university-bound students. In the time constraints of an English language course, effort can be focused on automatizing words below the productive vocabulary thresholds to prepare writers for first year undergraduate studies in English.

In reading, the 98% threshold of 8,000 word families (Laufer & Ravenhorst-Kalovski, 2010) is a useful goal for receptive vocabulary knowledge. In writing, the 98% threshold of 5,000 word families is a realistic target for productive vocabulary knowledge. These goals, and the frequency principle for learning vocabulary, are supported by the K8+ analysis in this study. Because of the infrequent use of lexical choices from beyond the 8,000 word frequency band of the BNC in novice university level writing, little instructional time and effort need be invested in learning these words before the first 5,000 (actively for writing) and the next 3,000 (passively for reading) word families are acquired. Once the most frequent 5,000 word families are learned productively and the overall most frequent 8,000 word families are learned receptively, the need for low frequency word choice is relatively rare, particularly in the early writing output of first year university students. Thus, these rare word choices can be safely mediated by students' underlying proficiency in their native tongues without overtaxing their mental capacity.

Productive thresholds also aid materials writers by indicating suitable lexical targets. In particular, textbooks can systematically focus on language below certain vocabulary thresholds to facilitate automaticity and the depth of vocabulary knowledge needed for productive use. The productive vocabulary thresholds thus further bring attention to the fact that vocabulary knowledge has to be developed beyond the receptive strands of reading and listening. Vocabulary should not be relegated only to the reading classroom. Rather, it has to also be developed in the writing classroom so that productive automaticity in the 5,000 most frequent word families in English is facilitated. Vocabulary is a vital part of the scope and sequence of an EAP writing course.

Limitations and Recommendations for Further Studies

Data were only available from students who gave permission for their EWT papers to be used anonymously for research purposes. As there were students who sat the EWT but did not provide permission, these data were lost to the researcher. Furthermore, the study is focused on one genre of academic writing, the essay test. Further studies in other genres found at the undergraduate level, such as laboratory reports and research papers, would round out an understanding of the productive vocabulary use of first year university

students. The study was also situated in one particular university setting, with a relatively homogenous group of undergraduate students. Further studies need to be carried out to investigate the productive vocabulary used at other institutions and by other groups of students. Finally, purposive sampling specifically chose papers for which at least one assessor had given a rating of *Satisfactory*, and most of the papers in the study were written by native English speakers. While earlier studies (Douglas, 2010) comparing essay tests written by students from English-speaking backgrounds and non-English-speaking backgrounds revealed significant differences between the two groups, a comparison of the LFPs of papers that had not passed the EWT and were written by native English speakers to the papers in the current study would be useful to see if there is a lexical difference between papers that had passed and had not passed when written by students from similar language backgrounds.

This study further provides a starting point for future cross-sectional or longitudinal studies focusing on the lexical breadth of knowledge in second, third, and fourth year undergraduate writing samples of university students. Studies that look at other levels of undergraduate writing will help to track the growth of productive vocabulary over a program of studies and inform supports for developing users of English at English medium institutions.

The current study focuses on breadth of vocabulary knowledge. An investigation into the depth of vocabulary knowledge in novice university level writing competence would contribute to elucidating the relationship between breadth and depth of vocabulary knowledge and the concomitant effects on assessment of writing competence. It would contribute to providing a complete picture of the lexical quality of entry-level post-secondary writing. By mapping out the productive lexical growth of undergraduate students, realistic and research based breadth and depth of knowledge vocabulary goals can be set for multilingual and developing users of English.

Finally, the current methodology lends itself to a diagnostic study of the productive vocabulary usage of developing users of English bound for undergraduate studies. The productive vocabulary output of students exiting from teacher-assessed EAP programs as well as students gaining entry to higher education through standardized English language proficiency tests can be compared to each other as well as to established lexical thresholds for novice university level writing competence. Coupled with this, an exploration of the relationship of productive vocabulary and eventual academic outcomes would provide information for curriculum planners on the impact of reaching vocabulary thresholds for students from non-English-speaking backgrounds.

References

Biber, D., & Conrad, S. (2001). Corpus-based research: Much more than bean counting. *TESOL Quarterly, 35*(2), 331-336.

Brynildssen, S. (2000). *Vocabulary's influence on successful writing*. (ERIC Digest D157). Bloomington, IN: ERIC Clearinghouse on Reading, English, and Communication. Retrieved from ERIC database. (ED446339)

Citizenship and Immigration Canada. (2013). Canada welcomes record number of international students in 2012 [Press release]. Retrieved from http://www.cic.gc.ca/english/department/media/releases/2013/2013-02-26.asp

Cobb, T. (2003). Analyzing late interlanguage with learner corpora: Quebec replications of three European studies. *The Canadian Modern Language Review, 59*(3), 393-423.

Cobb, T. (2012). Compleat lexical tutor v6.2. Retrieved from http://www.lextutor.ca

Conrad, S. (2005). Corpus linguistics and L2 teaching. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 393-409). Mahwah, NJ: L. Erlbaum Associates.

Corson, D. (1995). *Using English words*. Dordrecht, NL: Kluwer Academic Publishers.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238. Retrieved from http://www.jstor.org/stable/3587951

Coxhead, A. (2006). *Essentials of teaching academic vocabulary: English for academic success*. Boston, MA: Thomson Heinle.

Coxhead, A. (2011). The academic word list 10 years on: Research and teaching implications. *TESOL Quarterly, 45*(2), 355-362.

Coxhead, A., & Nation, P. (2001). The specialised vocabulary of English for academic purposes. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 252-267). Cambridge, UK: Cambridge University Press.

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In California State Department of Education (Ed.), *Schooling and language minority students: A theoretical framework* (pp. 3-49). Los Angeles, CA: Evaluation, Dissemination and Assessment Center, California State University, Los Angeles. Retrieved from http://www.eric.ed.gov/PDFS/ED249773.pdf

Douglas, S. R. (2010). *Non-native English speaking students at university: Lexical richness and academic success* (Unpublished doctoral dissertation). University of Calgary, Calgary, Canada. Retrieved from https://dspace.ucalgary.ca/bitstream/1880/48195/1/2010_Douglas.pdf

Effective Writing. (1993). *Detailed marking code*. Calgary, Canada: The Effective Writing Programme, University of Calgary. Retrieved from http://www.ucalgary.ca/writingsupport/markingcode

Effective Writing. (2003). *Assessors' guide for the Effective Writing Test*. Calgary, Canada: The Effective Writing Programme, University of Calgary.

Engber, C. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*(2), 139-155. Retrieved from http://www.sciencedirect.com/science/article/pii/1060374395900047

Gay, L. R., Mills, G. E., & Airasian, P. (2012). *Educational research: Competencies for analysis and applications* (10th ed.). Upper Saddle River, NJ: Pearson.

Gilmour, M. (2010, November). Student's can't write. *Macleans*. Retrieved from http://oncampus.macleans.ca/education/2010/11/19/students-cant-write

Grabe, W. (1985). Written discourse analysis. In R. B. Kaplan, A. d'Anglejan, J. R. Cowan, B. Kachru, G. R. Tucker, & H. Widdowson (Eds.), *Annual review of applied linguistics* (Vol. 5, pp. 101-123). New York, NY: Cambridge University Press.

Grayson, P. (2008). The experiences and outcomes of domestic and international students at four Canadian universities. *Higher Education Research and Development, 27*(3), 215-230.

Hart, B., & Risley, T. (2003). The early catastrophe: The 30 million word gap by age 3. *American Educator* (Spring 2003). Retrieved from http://www.aft.org/pdfs/americaneducator/spring2003/TheEarlyCatastrophe.pdf

Heatley, A., & Nation, P. (1994). Range [Computer program]. Wellington, New Zealand: Victoria University of Wellington. Retrieved from http://www.vuw.ac.nz/lals/

Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly, 37*(2), 275-301.

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*, 403-430.

Kelley, S. (2010, February). Texting, twitter contribute to students' poor grammar skills, prof says. *The Globe and Mail*. Retrieved from http://www.theglobeandmail.com/news/technology/texting-twitter-contributing-to-students-poor-grammar-skills-profs-say/article1452300

Krahn, H., & Taylor, A. (2005). Resilient teenagers: Explaining the high educational aspirations of visible minority youth in Canada. *Journal of International Migration and Integration, 6*(3-4), 405-434.

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316-323). Clevedon, UK: Multilingual Matters.

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal, 25*, 21-33. Retrieved from http://rel.sagepub.com/content/25/2/21.full.pdf+html

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 21-33.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15-30.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. New York, NY: Routledge.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27*(1), 57-86.

Ministry of Advanced Education, Innovation, and Technology. (2012). *British Columbia's international education strategy*. Retrieved from http://www.aved.gov.bc.ca/internationaleducation/forms/InternationalEducationStrategy_WEB.PDF

Morris, L., & Cobb, T. (2004). Vocabulary profiles are predictors of the academic performance of Teaching English as a Second Language trainees. *System, 32*(1), 75-87.

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

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

Nation, P. (2008). *Teaching vocabulary: Strategies and techniques*. Boston, MA: Heinle.

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage, and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, pedagogy* (pp. 6-19). New York, NY: Cambridge University Press. Retrieved from http://www.lextutor.ca/research/nation_waring_97.html

Raimes, A. (1983). Tradition and revolution in ESL teaching. *TESOL Quarterly, 17*(4), 535-552.

Raimes, A. (1985). What unskilled ESL students do as they write: A classroom study of composing. *TESOL Quarterly, 19*(2), 229-258.

Roessingh, H. (2006). BICS-CALP: An introduction for some, a review for others. *TESL Canada Journal, 23*(2), 91-96. Retrieved from http://www.teslcanadajournal.ca/index.php/tesl/article/viewFile/57/57

Roessingh, H. (2008). Variability in ESL outcomes: The influence of age on arrival and length of residence on achievement in high school. *TESL Canada Journal 26*(1), 87-107.

Roessingh, H., & Douglas, S. (2012). English language learners' transitional needs from high school to university: An exploratory study. *Journal of International Migration and Integration, 13*(3), 285-301.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Basingstoke, UK: Palgrave Macmillan.

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.

Smith, C. (2003). *Vocabulary's influence on successful writing: ERIC topical bibliography and commentary*. (ERIC Digest D157). Bloomington, IN: ERIC Clearinghouse on Reading, English, and Communication. Retrieved from ERIC database. (ED480633)

Spack, R. (1984). Invention strategies and the ESL college composition students. *TESOL Quarterly, 18*(4), 649-670. Retrieved from http://www.jstor.org/stable/3586581

Statistics Canada. (2012). *2011 Census of population: Linguistic characteristics of Canadians*. Retrieved from http://www.statcan.gc.ca/daily-quotidien/121024/dq121024a-eng.pdf

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

Wiggins, G., & McTighe, J. (2005). *Understanding by design, Expanded 2nd edition*. Upper Saddle River, NJ: Pearson Education.

Wilce, H. (2006, May). University students: They can't write, spell or present an argument. *The Independent*. Retrieved from http://www.independent.co.uk/news/education/higher/univesity-students-they-cant-write-spell-or-present-an-argument-479536.html

Zechmeister, E. B., D'Anna, C. A., Hall, J. W., Paus, C. H., & Smith, J. A. (1993). Metacognitive and other knowledge about the mental lexicon: Do we know how many words we know? *Applied Linguistics, 14*(2), 188-206. Retrieved from http://applij.oxfordjournals.org/content/14/2/188.full.pdf

Zechmeister, E. B., Chronis, A. M., Cull, W. L., D'Anna, C. A., & Healy, N. A. (1995). Growth of a functionally important lexicon. *Journal of Reading Behaviour, 27*(2), 201-212. Retrieved from http://jlr.sagepub.com/content/27/2/201.full.pdf