



**A Lexical Corpus Based Analysis of L2 Academic Vocabulary:
A Case Study**

Victor Khachan and Nahla Nola Bacha

Lebanese American University, Lebanon

Biodata

Victor Khachan is an Assistant Professor at the Lebanese American University where he has taught academic English for over fifteen years. He has published internationally and presented at international conferences. His research interests are corpus analysis, adult literacy, semiotics and academic writing.

Email: vkachan@lau.edu.lb

Nahla Nola Bacha is Professor at the Lebanese American University where she has taught academic English and administered for over twenty years. She has published internationally, worked with the Ministry of Education on the new textbooks, and presented at international conferences. Her research interests are in EAP/ESP, testing, and discourse analysis.

Email: nbacha@lau.edu.lb

Abstract

English proficiency tests are devised to assess student readiness for higher academic studies and to maintain institutional standards. Internationally, standardized tests are required by the majority of universities. However, with the increase in in-house English tests as an option in both L1 and L2 environments, the validity of measured parameters (such as writing) becomes an issue of urgent concern. Research has indicated that

investigation of active vocabulary in student writing is one way to address this question. Thus, this study profiles vocabulary in 103 essays (29,077 words) written by L1 Arabic speakers in a 'low' academic English proficiency course, as quantified by an in-house university English entrance exam (EEE) at an English medium university in Lebanon. Lexical vocabulary profiling (LVP) was carried out using Lextutor tools (available at www.lextutor.ca). Although the findings indicate consistent vocabulary use across the studied corpus, active vocabulary levels (K1, K2, and AWL lists) are not on par with the vocabulary profiles set by international standardized admissions tests as predictors of academic success. Recommendations are made to explicitly teach academic vocabulary.

Keywords: in-house English proficiency tests, EFL writing, academic word list, lexical vocabulary profiling, lexical corpus analysis, lexical recycle index

1. Introduction

English proficiency tests are devised to assess student readiness for academic studies and to maintain institutional standards, and thus continuous follow-ups on the validity of tests is important and is a regular feature of standardized tests such as the TOEFL and IELTS. Although standardized international English proficiency tests may dominate internationally, in-house English proficiency tests also have a strong presence at institutions throughout the world. These assessments normally evaluate discrete grammar, vocabulary, sentence structure, and paragraph or essay writing comparable to the international tests. However, with these in-house exams, their validity, especially for writing, is of concern to the institutions in which they are administered. One way to measure the validity of the writing produced on these tests is through investigating the vocabulary produced by test takers against relevant word lists using lexical profiling. The lexical profiling of available samples of the writing sections of the TOEFL and IELTS has associated cut-off scores with certain levels of vocabulary (Breeze, 2008; Dutton, 2006), which can aid investigation of student writing with in-house exams. Through vocabulary profiling techniques, the relationship between vocabulary level and writing quality can be assessed (Augstin Llach, 2005; Coxhead & Byrd, 2007; Horst & Collins, 2006).

Although validity studies are regularly carried out in international entrance exams, there are very few such studies investigating in-house exams in L2 contexts, and

thus this study evaluates the breadth and depth of active vocabulary produced on an in-house university English proficiency test developed at an English-medium university in Lebanon. Through lexical profiling, this study investigates the extent to which the tests measure vocabulary in writing and their ability to predict an active vocabulary profile of academic success as defined by international standardized tests. Pedagogically, the teaching/learning of vocabulary will be reflected upon in light of the findings in this study.

2. Review of Literature

Research has indicated vocabulary is an important indicator of academic writing proficiency with increased lexical variety and sophistication in the target language assessed as of higher quality (Morris, 2001; Morris & Cobb, 2004; Tschirner, 2007) and that developing vocabulary improves writing skills (Brynildssen, 2000).

In general, L2 written texts compared to those of native speakers are weaker and characterized by redundancy and simpler vocabulary (Bacha, 2002 & 2005; Morris & Cobb, 2004) and considered of minimal proficiency for university academic work (Gilquin, Granger, & Paquot, 2007; Mukattash, 2003; Pool, 2003). In order to help students widen their vocabulary repertoire, some comparative research has examined L2 academic writing texts (used interchangeably with writing or essays in this study) and university academic reading. This has resulted in the production of some vocabulary texts. These texts are helpful to students in preparing for English entrance exams and/or developing their academic vocabulary for more effective academic writing (Praniskas, 1972 and Yorkey, 1981, as cited in Coxhead, 2000). Other similar studies have also investigated the academic vocabulary required for university work (Breeze, 2008; Coxhead & Nation, 2001; Dutton, 2006). Recent studies in corpora in English for Academic Purposes (EAP) pedagogy have focused on analyzing vocabulary in both student and professional texts, predicting academic success, determining student proficiency levels, assessing development and progress, and using the results in materials design for both courses and EFL textbooks (Coxhead & Byrd, 2007; Krishnamurthy & Kosem, 2007). Furthermore, studies in disciplines such as medicine, engineering, anatomy, professional air tourism, and applied linguistics have shown the value of corpus lexical analysis in determining the words in professional texts and thus the vocabulary

necessary for students to write effectively (Chen & Ge, 2007; Wang, Liang, & Ge 2008; Mudrdaya, 2006). In fact, the academic performance of non-native speakers could be obtained along with other text and non-text indicators through assessing the vocabulary in their writing. To highlight some of this research, Lee and Muncie (2006) investigated the effects of intervention of explicit vocabulary strategies and integration of language skills with high school ESL learner use of vocabulary in writing, showing an increase in higher target level vocabulary above the 1,000-2,000 word level, improving the lexical frequency profile of their writing. A few studies on the growth of learner lexicons (Lenko-Szymanska, 2000) have shown the validity of using the lexical frequency profiles to measure vocabulary output and the pedagogical implications of frequency lists and lexical analysis in corpus studies for L2 writing.

In justifying our evaluation of the writing quality through lexical assessment measures, we refer to research that has further indicated word lists such as the Academic Word List (AWL) are of value in assessing the vocabulary level of students' written texts which help determine student proficiency levels and give insights to developing lexical levels and thus facilitate more effective writing (Coxhead, 2000). Coxhead's (2000) AWL has provided researchers and practitioners with an assessment tool that accounts for the type of active vocabulary in students' texts against a criteria of 570 word families (10 sublists) needed for effective writing and thus successful academic study. The General Word List (GWL), the University Word List (UWL), and the Thorndike Word Lists have also provided researchers with the most frequent word levels required for students to understand and write academic texts. Nation (1990) argues that the threshold needed for reading in English depends on learners knowing the first 2,000 most frequent words, but in university settings those words must be augmented with more specific academic language, as supplied by lists such as the AWL. Some researchers quantify the assessment of required vocabulary in writing for different academic levels. Tschirner's (2007) study outlines the threshold levels for different grades in pre-tertiary institutions and the levels of passive vocabulary (i.e., those words that students can recognize and understand in reading texts) and active vocabulary (i.e., those words that students can produce in their writing) indicating that in grades 11 and 12 (almost equivalent to the Remedial course in this study) productive (or active) words are 4,000. Tschirner (2007) further reports on studies in which a minimum estimated vocabulary size needed for

academic purposes is approximately 5,000 words for authentic texts and a range from 5,000 to 10,000 for university textbooks. Nation (2006) argues that at least 97% (8,000-9,000 word families) of the vocabulary of a text needs to be known to gain adequate understanding of the text. Nation (2006) reports that native and non-native learners gain these levels through not only reading a lot but also more efficiently, coupled with learning lists such as the AWL. Tschirner (2007) recommends schools and universities give more direct vocabulary learning even for advanced L2 students since there are significant differences between native speakers in vocabulary development through guessing and retention strategies (Laufer, 2003 in Tschirner, 2007). Cobb & Horst (2001) report that if the first 2,000 words are known (referred to as K1 and K2 hereafter) and the 570 AWL word families, then the learner knows about 90% of the vocabulary they should meet in any academic text and support this by referring the reader to computer text analyses. Nation & Beglar (2007) report that learners need 98% of the vocabulary of written and/or oral texts to comprehend texts unassisted (Hu & Nation, 2000, as cited in Nation & Beglar, 2007). They emphasize that the three reasons for assessing the vocabulary is to determine how close learners are to the minimum requirement, to monitor learner vocabulary development, and to make comparisons of the level and rate of development with that of native speakers. Horst (2005) adds that L2 learners widen their vocabulary through a great deal of extensive reading which can be assessed through frequency profiling. Since the AWL is used as a lexical measure in this study, some additional explanation about it is necessary. There has been some debate concerning the AWL as a valid lexical measure and, therefore, its efficacy for assessing the quality and level of L2 writing (see Hyland & Tse, 2007). Hyland & Tse (2007) contest the existence of an academic vocabulary arguing that specific vocabulary lists for the various disciplines makes more sense, and thus the AWL is inadequate. However, Eldridge (2008) finds value in the AWL, pointing out:

students are more involved in general academic English in taking the international exams IELTS and TOEFL and the general environment involving instructions and communicating in an academic environment makes acquiring a general academic vocabulary for interdisciplinary communication important. (p. 110)

Eldridge (2008) debates whether students would be better off in acquiring a “universal literacy or multiple literacies depending upon the context” (p.110). Taking the foregoing into account, Eldridge (2008) reports “Coxhead’s flawed AWL may continue for a while to be of more practical service than the specialized approach suggested by Hyland and Tse (p.111). Having said all of this, research using the AWL vocabulary profile has indicated the levels of the vocabulary in texts which provide guidelines for university admissions (Breeze, 2008; Dutton, 2006), which in turn give insights into needed vocabulary development in L2 written texts (Coxhead, 2000).

3.1 Aim of Study

Through corpus analysis, this study aims to investigate the extent to which lexical profiling measures vocabulary in writing and their ability to predict an active vocabulary profile of academic success defined by international standardized tests. The teaching/learning of vocabulary is reflected upon in light of the findings in the study. In this way, the study contributes to evaluating active vocabulary as a determiner of writing quality.

3.2 Research Questions

1. Is the writing of students with EEE 500-549 an acceptable English proficiency threshold required for university study? In other words, is the entry writing proficiency of the in-house entrance exam on par with those of the international standardized tests of the TOEFL, SAT, and IELTS?
2. Is active vocabulary a determining factor in overall writing assessment? In other words, to what extent does the vocabulary, as operationalized by the AWL (Coxhead, 2000), compare to that of the international standardized essay tests of the TOEFL and IELTS?

4. Method

4.1 Student background

Although Lebanese English-medium universities accept international standardized English proficiency tests, the number of students opting for such international tests remains restricted to overseas students. At the time of the study, the local population of

students was channeled mostly via the in-house English proficiency test which, to a great extent, resembles the paper-based TOEFL (discrete item assessment of grammar, reading/comprehension, and essay writing). Students were admitted and placed in the EFL courses based on cut-off scores. To our knowledge, it was run by experts and was the sole responsibility of the University Testing Office. The EFL Program was not involved in writing, administration, or correction of the exam.

The feeder schools to the university are of two types: private and public. Private schools are mainly sponsored by the French or Americans, with French as the medium of instruction in the former and English in the latter. Students are referred to as either French or English educated. The other feeder schools are public governmental schools in which French is the language of instruction. Most students are L1 Arabic speakers. According to statistics from the Lebanese Ministry of Education (2009), 62.5% of all Lebanese schools offered French as a second language in the school year in 1999-2000, which decreased to 55.8% in 2005-2006, and schools in which English was offered increased from 19.7% to 21.6%. It is clear universities need to address the issue of valid English entrance exams as more students opt to enter English medium universities.

4.2 Data Collection

The present study was carried out as a diagnostic essay test given in the first week of classes to students who were enrolled in the first English course, ENG009, a remedial non-credit English course with a three-credit teaching load. It was not possible to obtain the writing from the in-house entrance exam (used interchangeably with EEE hereafter), but these essays approximated those of the writing section on the in-house entrance exam (see below for a description). In this context, it was a purposeful sample in that all the students registered in the course took part. Participant ages were between 17-18 with Arabic as their first language and French and/or English their second. Students were enrolled in different disciplines at the university taking four courses in addition to remedial English.

To approximate the EEE testing environment, students were briefed on the seriousness and importance of the diagnostic writing activity for teachers to pinpoint students' individual weaknesses so they could assign bridging tasks/activities. In this context, the data-collection framework approximates the testing environment of the

writing component of an EEE writing test. To ensure time proximity and similar writing performance, it was important to limit the time between EEE results and data collection. Similar to the EEE writing environment, the diagnostic test/essay used in this study was given in a 45-minute sitting without external help (such as dictionaries); and students were assigned a typical EEE topic which dealt with *parents and teenagers* where students could draw upon their own experiences in answering.

The academic writing corpus collected for this study comprised the essay writing of 103 first-year university students. In terms of English proficiency, students admitted had to secure a minimum EEE score of 500. Students with EEE scores between 500-549 points (or SAT writing section 380 points, internet-based TOEFL (iBT) 80-90 points, or IELTS 6.5 points) were automatically channeled into ENG009 and those with EEE scores between 550-599 (iBT 91-100) were exempted from English 009 and admitted into a higher course.

4.3 Procedure

The 103 diagnostic essays were all photocopied and transcribed; the typists ensured the correction of spelling mistakes, except those denoting grammatical errors, as required for digital analysis of the vocabulary. Essays were compiled into individual files according to five course sections. Finally, the five sections were combined to create this study's ENG009 Learner Corpus (see Table 1), a mini-corpus topic-based mini corpus of student writing (similar to that of Schmitt & Schmitt, 2007).

The text files were individually uploaded to the Web Vocabulary Profiler (Web VP, version 1.5 also known as VocabProfile) as part of the Compleat Lexical Tutor, available at www.lextutor.ca (Cobb, 1999-2009). Data analysis was carried out on three Lextutor fronts: 1) Text LexCompare, 2) VocabProfile and 3) British National Corpus (BNC) as a reference corpus.

The upload process resulted in the following output findings, representing the 5 ENG009 sections and the ENG009 Learner Corpus (the 5 sections' sub-corpora combined).

5. Data analysis and discussion

Table 1 indicates the average length of each of the produced essays corresponds with EEE requirements as per EEE writing instructions: a minimum of 1½ handwritten pages (averaging 250-300 words).

Table 1. ENG009 Learner Corpus: Participants, corpus size, and average essay length

	No. of Subjects	Words/Sub-corpus	Words/subject
Sub-corpus 1	15	4116	274.4
Sub-corpus 2	22	5716	259.81
Sub-corpus 3	25	8033	321.32
Sub-corpus 4	21	4788	228
Sub corpus 5	20	6419	320.95

Although these averaged essay length figures do not necessarily mirror the actual length of individual essays, this approximation confirms, to a great extent, the nature of the ENG009 Learner Corpus compiled for this study. In addition, the average essay length is in line with most of the studies that have employed VocabProfile in the past (see Dutton, 2006; Horst & Collins, 2006; Laufer & Nation, 1995).

Lexical recycling/repetition

In general, lexical recycling/repetition through text comparison is devised to calculate the range of repeated words (and unrepeated/new) across a variety of texts. From a practical learning perspective, text comparison programs such as TextLexCompare (available at www.lextutor.ca/text_lex_compare) are engineered to trace vocabulary learning/acquisition opportunities from one text to another. By calculating the recycle index, or the number of recycled words divided by total words in the new text, TextLexCompare highlights lexical prominence between an old and new text, and thus determining learning possibilities/opportunities based on the frequency of encountered words.

However, in this study, lexical recycling was used as a tool to measure similarity among the 5 sub-corpora; the higher the similarity rate (repetition/recycling), the closer vocabulary proximity/choice is across all sub-corpora. That is, a high rate of repeated/recycled words from one sub-corpus to another is a direct indication that

ENG009 students have similar productive/active vocabulary thresholds. As shown in Table 2, lexical recycling varies from 93.67% to 95.06% across all sub-corpora. The lowest recycling rate shows 93.67% of the words in sub-corpus 3 are repeated in sub-corpora 1, 2, 4, and 5. On the other hand, sub-corpus 2 has the highest rate of repeated words (95.06%) in comparison with the remaining 4 sub-corpora.

Table 2. ENG009 Learner Corpus: participants, corpus size and average essay length

	No. of Subjects	Words/Sub-corpus	Words/subject
Sub-corpus 1	15	4116	274.4
Sub-corpus 2	22	5716	259.81
Sub-corpus 3	25	8033	321.32
Sub-corpus 4	21	4788	228
Sub-corpus 5	20	6419	320.95

Considering Cobb's (2007) first indication that a rate of 70% vocabulary repetition/recycling can be associated with "related or sequential texts by the same author" (p. 48). The high recycle index rates across the ENG009 Learner Corpus are an indication of writing consistency and empirical support favoring the EEE cut-off scores (500-549), but not necessarily a key point in validating those scores against international standardized English proficiency tests. Recycle indices found across the 5 sub-corpora indicate fresh ENG009 students/EEE 500-549 performers have similar pools of vocabulary.

Yet questions remain as to whether standardized international tests would evaluate the EEE 500-549 students as the lowest acceptable English proficiency threshold required for university study and whether active vocabulary represents a determining factor in overall writing assessment. To put these questions into perspective, EEE 500-549 scores must be looked at vis-à-vis matching TOEFL iBT and IELTS scores recognized at the university. The EEE 500-549 equivalent TOEFL (iBT), IELTS and SAT are 80-90 points, 6.5 points and 380 points (writing section), respectively. IELTS 6.5 performance band, corresponding to the lowest acceptable EEE performance (500-549), is defined by the Common European Framework (2001) as "an advanced level of competence suitable for more complex work and study tasks" (Council of Europe, 2001,

p. 2, quoted in Shaw & Weir, 2007, p.159). The problem, then, is not a matter of consistency with the IELTS cut-off score band, but is rather related to the IELTS assessment criteria, mainly those surrounding vocabulary distribution. In his lexical investigation of IELTS speaking tests, Read (2005) reports on performance bands 4-8. The size of Read's corpus, mainly band 6 (18,493 words) and band 7 (21,865 words), the closest to EEE 500-549, is within proximity to this study's ENG009 learner corpus.

Read (2005) concludes that vocabulary sophistication and frequency are associated with band performance, band 8 being the most sophisticated whereas band 4 the least. Applying Read's findings to the present study, it is found that band 4, which is 2.5 performance points lower than the university in this study's English proficiency admission requirement, is much more sophisticated than that of the average performance in the ENG009 corpus (see Tables 3 and 4).

To conclude, the recycle index results in this paper demonstrate international standardized test findings and expectations are drastically more demanding than those found in the EEE.

Table 3: ENG009 Learner Corpus: participants, corpus size and average essay length

	No. of Subjects	Words/Sub-corpus	Words/subject
Sub-corpus 1	15	4116	274.4
Sub-corpus 2	22	5716	259.81
Sub-corpus 3	25	8033	321.32
Sub-corpus 4	21	4788	228
Sub-corpus 5	20	6419	320.95
ENG009 Learner Corpus	103	29077	282.3

Table 4: ENG009 Sub-corpora TextLexCompare Output

Old files	New file	Tokens Recycle Index
Sub-corpora 2-5	Sub-corpus 1	94.57 %
Sub-corpora 1, 3, 4 & 5	Sub-corpus 2	95.06%
Sub-corpora 1, 2, 4 & 5	Sub-corpus 3	93.67%
Sub-corpora 1, 2, 3 & 5	Sub-corpus 4	95.04 %

Sub-corpora 1-4 & 5	Sub-corpus 5	94.76%
---------------------	--------------	--------

K1 vocabulary distribution

In light of Coxhead's (2000) vocabulary distribution (Table 5) and Nation's (2001) analysis of the Brown corpus (Table 6), Table 3 reveals that the K1 percentile distribution across all ENG009 sub-corpora is higher than expected (percentages varied from 80.68% to 82.68%). The study's corpus vocabulary profiling reveals a consistently high dependency of ENG009 students on the first 1000 most frequent words in English.

Table 5. Coxhead's (2000) percent distribution of vocabulary in academic texts

Sub-corpus	AWL	1 st 1000 words (GSL)	2 nd 1000 words (GSL)	Total
<i>Arts</i>	9.3	73	4.4	86.7
<i>Commerce</i>	12	71.6	5.2	88.8
<i>Law</i>	9.4	75	4.1	88.5
<i>Science</i>	9.1	65.7	5	79.8

Table 6. Percent of the most frequent word families in an average text as revealed in the Brown corpus (taken from Nation, 2001)

Word Families	Percent (%) of Words in Average Text
10	23.7
1000	72
2000	79.7
3000	84
4000	86.7
5000	88.6
6000	89.9

K1 vocabulary allocations in Coxhead's (2000) analysis varied according to the type of academic sub-corpus. K1 distribution in Coxhead's arts sub-corpus, the closest to the present study, accounts for 73% of the texts (see Table 5). However, in the context of the British National Corpus (BNC) as a reference corpus, ENG009 Learner Corpus K1

vocabulary distribution is quite noticeable. The ENG009 corpus average distribution shows a high 88.86% whereas figures for sub-corpora 1 to 5 are 89.48%, 89.22%, 88.94%, 87.10%, 89.36%, respectively (see Table 7).

Table 7. ENG009 Learner Corpus VocabProfile Output Based on BNC as a Reference Corpus

BNC	ENG009 Corpus %	Sub- corpus 1	Sub- corpus 2	Sub- corpus 3	Sub- corpus 4	Sub- corpus 5
K1	88.86	89.48	89.22	88.94	87.10	89.36
K2	5.16	4.65	4.71	5.07	6.46	5.04
K1-K2	94.02	93.13	93.93	94.01	93.56	94.4
K3	3.86	3.52	4.28	3.48	4.47	3.71
K1-K3	97.88	96.65	98.21	97.49	98.04	98.11
K4	0.83	1.12	0.61	0.85	0.83	0.82

When learner performance on the international standardized IELTS and in light of the context of Nation's (2001) K1 vocabulary distribution of the Brown corpus (Table 6) was considered, the ENG009 corpus K1 vocabulary distribution findings remain inflated (Table 7). The learner corpus has a frequency average of 82.21% with a consistent range of 80.68% to 82.68%; whereas in the IELTS writing sample vocabulary, frequencies in bands 6 and 7 (6.5 being equivalent to the EEE 500 cut off score for university entrance) indicate 59.5% and 54.6% respectively (Read, 2005, p.14). This indicates that the learner corpus contains an inflated percentage of high frequency words, when compared to the IELTS writing sample (Read, 2005).

K2 vocabulary distribution

The VocabProfile output in Table 3 shows K2 distributions of 6.16% to 6.75% across all sub-corpora, figures close to the K2 vocabulary distribution in the Brown corpus of 7.7% (Nation, 2001; see Tables 3 and 6). Additionally, analysis of the ENG009 corpus and individual sub-corpora against the British National Corpus (BNC) reveals K2 distribution close to Coxhead's K2. Table 7 shows the BNC K2 distribution for the ENG009 corpus is 5.16% whereas sub-corpora 1 to 5 vary from 4.65% to 6.46%, percentages that are

comparable with Coxhead's academic corpus (4.4%, 5.2%, 4.1% and 5%) in arts, commerce, law, and science respectively (see Table 5).

Nevertheless, when the IELTS cut-off score of 6.5 (equivalent to the EEE500-549) as a university entrance requirement is considered, the learner essays then become poor by comparison as the IELTS emphasizes a K2 frequency of 14.9% to 15.1% (Read, 2005) which the learner texts do not exhibit.

To conclude these sections, the ENG009 corpus indicates a high vocabulary frequency distribution of 94.02% with a range of 93.13% to 94.4%, much higher than Coxhead's (2000) highest K1-K2 vocabulary distribution of 79.1%. It is expected that the lower the K1-K2 frequency, the more academic the text is (Dutton, 2006). The ENG009 corpus thus indicates less sophisticated vocabulary and a higher frequency of more common words (Read, 2005), reaffirmed when compared to the IELTS band 6 K1-K2 word frequency average of 74.4% (Read, 2005).

Academic vocabulary distribution

Table 3 indicates the active use of academic vocabulary, as defined by Coxhead's (2000) AWL list, is minimal and below academic writing requirements or expectations. Academic texts on average, according to Coxhead (2000), manifest an AWL distribution of 10%. Except for sub-corpus 4, AWL distribution in this study does not cross the boundary of 3.82%, about three times below the academic vocabulary ceiling set by Coxhead (2000) and Read (2005). However, when this low rate distribution is considered in light of the AWL 10 sublists (i.e., 570 word families), ENG009 Learner Corpus pinpoints an academic vocabulary across the 10 sublists (see Table 8).

Table 8. Presence of academic vocabulary in ENG009 Learner Corpus across Coxhead's (2000) AWL sublists

AWL Sublists	Presence of AWL families in ENG009 corpus (out of 570 families)
Sublist 1	Analyse approach assume authority available benefit concept consist create derive economy environment factor finance identify income individual involve issue major method occur percent period principle proceed process require research respond role section sector significant similar source specific theory (38 families of 60) [63.3%]

Sublist 2	achieve affect appropriate aspect category community compute conclude consequent culture design final focus impact injure normal obtain positive previous primary range regulate restrict secure seek site text tradition transfer (29 families out of 60) [48.3%]
Sublist 3	circumstance constant contribute corporate emphasis illustrate imply instance interact negate outcome physical react sex shift sufficient technology (17 families out of 60) [28.3%]
Sublist 4	Access adequate apparent attitude attribute commit communicate concentrate contrast cycle goal impose integrate job label mechanism obvious phase predict principal project resolve statistic stress sum (25 families out of 60) [41.6%]
Sublist 5	academy aware capacity challenge conflict contact draft energy evolve expose facilitate generate generation image licence medical mental network psychology reject stable style trend whereas (24 out of 60) [40%]
Sublist 6	acknowledge attach bond brief capable discriminate diverse furthermore ignorant intelligence minimum neutral nevertheless precede reveal transform transport (17 families out of 60) [28.3%]
Sublist 7	adapt adult comprehensive confirm convert couple decade definite globe grade ideology innovate isolate media mode phenomenon prohibit reverse somewhat survive topic transmit ultimate visible (24 families out of 60) [40%]
Sublist 8	contradict deviate drama eventual exhibit highlight induce intense manipulate plus radical random tense (13 families out of 60) [22.3%]
Sublist 9	diminish duration ethic mature military mutual norm relax revolution vision (10 families out of 60) [16.6%]
Sublist 10	collapse convince persist whereby (4 families out of 30) [13.3%]

Coxhead (2000) highlights the prominence of her 10 sublists, associating sublist 1 (the most frequent 60 words) with 3.6% coverage of the words in an academic text. Academic vocabulary distribution in sublists 2 (2nd most frequent), 3 (3rd most frequent) and 4 (4th most frequent) cover 1.8% , 1.2%, and 0.9%, respectively. As table 9 indicates, a mastery of AWL sublists 1-4, according to Coxhead (2000), ensures the coverage of 7.5% of the 10 % available in academic texts. The ENG009 corpus indicates sublist 1 coverage of 63.3% whereas the coverage of sublists 2, 3, and 4 is distributed as 48.3%, 28.3% and 41.6% respectively.

Table 9. Coxhead's (2000) distribution of academic vocabulary across AWL sublists

Sublist	Items	Coverage of Academic Corpus (%)	Cumulative Average (%)
1	60	3.6	3.6
2	60	1.8	5.4
3	60	1.2	6.6
4	60	0.9	7.5
5	60	0.8	8.3
6	60	0.6	8.9
7	60	0.5	9.4
8	60	0.3	9.7
9	60	0.2	9.9
10	30	0.1	10

The average 3.8% AWL in the ENG009 corpus and the consistent range from 3.52% to 4.47% across the five sub-corpora based on the Brown corpus indicate a very low frequency of academic words in the texts when compared to Coxhead's (2000) target of 10% and even to Nation's (1990) lower target of 8%. When compared to Read's (2005) study, the IELTS bands 6 and 7 are 10% and 9.4% respectively, with the TWE corpus at the 9.49% academic word frequency level (Breeze, 2008) and the learner's corpus in Dutton's (2006) study 7.17%. Table 9 indicates that when the ten sublists of the AWL were examined in the ENG009 corpus, 201 families out of the 570 are present in the corpus as a whole, which shows an adequate representation of the different types of words in the AWL. The implication is that teachers should draw on the words in the AWL and explicitly teach the words and not expect students to pick them up through context or their own reading. Coxhead (2000) states:

Even though Sublists 5-10 add little to the overall coverage of the AWL, they are worth including, as these less frequent items occur in a wide range of texts and are unlikely to be acquired incidentally through reading (p.228).

These results have strong implications for teaching and learning academic vocabulary in students' first years of study, a call to action endorsed by Breeze (2008) and Coxhead & Byrd (2007).

6. Implications and recommendations for future research

This study investigated the frequency of vocabulary in L1 Arabic students' writing in an in-house English entrance exam. Specifically, the study investigated to what extent the in-house exam compared to the international tests as measured through vocabulary profiling and whether it was on par with those tests. On both accounts, the main results indicate the academic vocabulary in the in-house entrance writing exam is not as challenging as the international writing exams, indicating lower writing quality. However, only individual words were examined, not the words in context or cohesion and coherence, a limitation of the study. Qualitative analysis of the students' writing is needed to obtain a more comprehensive understanding of the lexis in the students' texts. Furthermore, the essay exams of the in-house test and those of the international exams were written on different topics and had different populations and numbers which may influence the results. However, since the aim of the study was to investigate the general quality of the writing according to a general AWL and not specific lexical items per se, this comparative study design was considered satisfactory.

The results have implications for explicit teaching and learning of academic vocabulary in the students' first year of study at the university. In fact, through corpus studies such as the present one, much can be learned from students' academic writing (see Coxhead, 2010). Brynildssen (2000) reports vocabulary development should be part of teaching and learning in classrooms, as words form the basis of any writing. Based on the results in this study and research in the field, some recommendations are made below for vocabulary teaching and learning in the remedial academic English classes at the university:

- Selecting interesting readings from various sources rich with vocabulary
- Involving students in group vocabulary activities
- Keeping journals in which students use the new vocabulary
- Having a class newsletter in which different types of writing are included

- Giving time for writing assignments in and out of class
- Conferencing with teachers
- Assessing the level of student vocabulary, and thus writing proficiency, through computer-based tools such as VocabProfile

Although this study focuses on one remedial course, it confirms the value of the AWL in L2 writing research. Additionally, we believe a new research culture has been established in the EFL Program in this study which should be expanded. Since words are the building blocks of texts, it is these blocks we need to increase, relevant to students' studies. Future research is needed to investigate any lexical development over time and the extent of the development after explicit program instruction has been given. A comparison, in addition, with the level of vocabulary in the required textbooks would raise awareness of the specific vocabulary students need. The findings of the present study imply institutions of higher education must be aware that developing and administering in-house English proficiency exams may eventually benefit many universities, for such exams can be tailored to the needs and expectations of the teaching body. In practice, however, the legacy of in-house English proficiency exams has proven otherwise and often the validity of these exams is questionable. The development of an English proficiency exam requires an academic think-tank, with high priority given to the academic constituencies and experts in testing and evaluation. The role of the English teaching body, on the other hand, must be a complementary one, assisting in the realization of a language culture that develops teaching and learning strategies to link pre-university language proficiency to post-university market demands. All in all, once a university decides to develop and administer in-house English proficiency exams, it becomes that university's responsibility to continuously validate this tool, and here we offer one means of accomplishing such validation.

References

Augustin Llach, M.P. (2005). The relationship of lexical error and their types to the quality of ESL compositions: an empirical study. *Porta dLinguarum*. 3, 45-57.

- Bacha, N. N. (2002) Developing learners' academic writing skills in higher education: A study for educational reform. *Language and Education International Journal*. 16(3): 161-177.
- Bacha, N. N. (2005). Academic vocabulary: A corpus analysis approach. *International Journal of Arabic-English Studies (IJAES)*. 6(1): 123-146. Librarie du Liban Publishers.
- Breeze, R. (2008). Researching simplicity and sophistication in student writing. *International Journal of English Studies*, 8(1), 51-66.
- Brynildssen, S. (2000). Vocabulary's influence on successful writing. ERIC Clearinghouse on Reading, English, and Communication Digest #157 retrieved February 21, 2011 from <http://www.indiana.edu/~reading/ieo/digests/d157.html>
- Chen, Q. & Ge, G. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes*. 26(1), 103-120.
- Cobb, T. (2008). Commentary: response to Mcquillan and Krashen. *Language Learning & Technology*. 12(1), 109-114.
- Cobb, T. (2007). Why and how to use frequency lists to learn words. www.wordhacker.com, retrieved March 16, 2009.
- Cobb, T. (1999-2009). Compleat Lexical Tutor, available at www.lextutor.ca.
- Cobb, T., & Horst, M. (2001). Reading academic English: Carrying learners across the lexical threshold. In J. Flowerdew & A. Peacock (Eds.), *The English for Academic Purposes Curriculum* (pp. 315-329). Cambridge: Cambridge University Press.
- Coxhead, A. (2010). What can corpora tell us about English for Academic Purposes? In A. O'Keefe; M. McCarthy (Eds.), *The Routledge handbook of Corpus Linguistics*, (pp. 458-470). Oxford: Routledge.
- Coxhead, A. (2000) A new academic word list (Electronic version). *TESOL Quarterly*. 34(2), 213-238.
- Coxhead, A. & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing*, 16, 129-147.
- Coxhead, A. & Nation, P. (2001). The specialized vocabulary of English for academic

- purposes. In John Flowerdew & Matthew Peacock (eds.) (pp. 253-267) *Research Perspectives on English for Academic Purposes*. Cambridge: Cambridge University Press.
- Dutton, M. (2006). Word lists tagging for teachers: Investigating the correlation of the general service list and academic word list to IELTS task 2 writing. *19th Annual EA Education Conference*, Perth, Australia.
- Eldridge, J. (2008). No, there isn't an 'academic vocabulary, but ...' A reader responds to K. Hyland and P. Tse's Is there an 'academic vocabulary'? *TESOL Quarterly* 42(1): 109-113.
- Engber, C.A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*. 4(2), 139-155.
- Gilquin, G., Granger, S. & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes*, 6, 319-335.
- Green, A. (2007). IELTS washback in context : preparation for academic writing in higher education. [Studies in language testing ; 25](#); Cambridge University Press: Cambridge
- Horst, M. (2005). Learning L2 Vocabulary through extensive reading; a measurement study. *The Canadian modern language review*. 61(3), 355-382.
- Horst, M. & Collins, L. (2006). From *faible* to strong: How does their vocabulary grow? *The Canadian Modern Language Review*, 63(1), 83-106.
- Hyland, K., & Tse, P. (2007). Is there an "Academic Vocabulary"? *TESOL Quarterly*, 41(2), 235-253
- Krishnamurthy, .R. & Kosem, I. (2007). Issues in creating a corpus for EAP pedagogy and research. *Journal of English for Academic Purposes*. 6, 356-373.
- Laufer, B. & Nation, P. (2001). Passive vocabulary size and speed of meaning recognition. *Eurosla Year Book*. 1, 7-28.
- Laufer, B., & Nation, P. (1995). Vocabulary size: lexical richness in L2 written production, *Applied Linguistics*, 16, 307-322.
- Lee, S. H. & Muncie, J. (2006). From receptive to productive: Improving ESL learners' use of vocabulary in a postreading composition task. *TESOL Quarterly*, 40(2), 295-320 (26).
- Lenko-Szymanska, A. (2000). How to trace the growth in learners' active vocabulary? A

- corpus-based study. *Language and Computers, Teaching and Learning by Doing Corpus Analysis*. In Kettemann, B. & Marko, G. (Eds.). *Proceedings of the Fourth International Conference on Teaching and Language Corpora*, Graz. July 19-24, (pp. 217-230 (14)).
- Morris, L. (2001, Nov). The use of vocabulary profiles in predicting the academic and pedagogic performance of TESL trainees. *International Language in Education Conference*. Hong Kong, China.
- Morris, L. & Cobb, T. (2004). Vocabulary profiles as predictors of the academic performance of teaching English as a Second Language trainees. *System*, 32, 75-87.
- Mudrdaya, O. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purposes*. 25, 235-256.
- Mukattash, L. (2003). Towards a new methodology for teaching English to Arab learners (TEAL). *International Journal of Arabic-English Studies*, 4, 211-234.
- Nation, I.S.P. (1990). *Teaching and Learning Vocabulary*. New York: Newbury House.
- Nation, I.S.P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.
- Nation, I.S.P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language*. 61, 59-82.
- Nation, P. & Beglar, D. (2007). A vocabulary size test. *The Language teacher*. 31(7), 9-13.
- Poole, G. (2003). Assessing Japan's institutional entrance requirements. *Asian EFL Journal*. (5)1, www.asian-efl-journal.com/march03.sub5aphp, retrieved March 16, 2009.
- Read, J. (2005). Applying lexical statistics to the IELTS speaking test. *Research Notes*. 20, 12-16.
- Schmitt, D. & Schmitt, N. (2007). Overcoming the gap in exposure to text: using a reading corpus in an academic writing class. TESOL 2007 conference.
- Shaw, S. D. & Weir, C.J. (2007). *Examining writing: research and practice in assessing second language writing*. Cambridge: UCLES/Cambridge University Press
- Tschirner, E. (2007). Breadth of vocabulary and advanced English study: An empirical investigation. *Electronic Journal of Foreign Language Teaching*. 1(1), 27-39.

Wang, J. Liang, S.I. & Ge, G-C. (2008). Establishment of a medial academic world list.
English for Specific Purposes. 27, 442-458.