Journal of English for Academic Purposes 21 (2016) 121–132

Contents lists available at ScienceDirect

Journal of English for Academic Purposes

journal homepage: www.elsevier.com/locate/jeap

Forum

The relationship between academic vocabulary coverage and scores on a standardized English proficiency test☆

T. Sima Paribakht [a, *], Stuart Webb [b]

[a] Official Languages and Bilingualism Institute, University of Ottawa, 70 Laurier Avenue East, # 123, Ottawa, Ontario, K1N 6N5, Canada
[b] Faculty of Education, Western University, 1137 Western Road, # 1026 London, Ontario, N6G 1G7, Canada

ARTICLE INFO

Article history:
Received 8 January 2015
Received in revised form 14 May 2015
Accepted 27 May 2015
Available online 9 June 2015

Keywords:
Vocabulary
Lexical profile
Academic word list (AWL)
Lexical coverage

ABSTRACT

This study examined the occurrence of items from the Academic Word List (AWL) (Coxhead, 2000) in reading, listening and cloze texts in 12 versions of an English proficiency test used for admission purposes at Canadian universities. The relationship between the coverage of AWL in the texts and the corresponding scores of 6380 test takers was also analyzed. The results indicated that AWL coverage in the passages was consistently present and substantial although below the established level of coverage found in academic texts. Furthermore, AWL coverage had no correlation with the overall reading and listening comprehension scores, and there was a small, negative correlation with cloze test scores. The findings are discussed in light of lexical specifications of language proficiency tests used for university entry purposes.

1. Background

1.1. Corpus-based research in applied linguistics

The use of computer tools in applied linguistics research has allowed for compilation of large electronic linguistic corpora and their analysis for a variety of purposes over the past few decades (e.g., The British National Corpus (BNC) with over 100 million words, The (COBUILD) Bank of English Corpus with over 300 million words, the Cambridge International Corpus (CIC) with over 100 million words, The Corpus of Contemporary American English (COCA) with 450 million words). Access to such large data bases combined with the speed of data processing and analysis afforded by computer software, provide the grounds for examination of the frequency of occurrence of various language elements/features and the patterns of their use in authentic language. This has opened up new areas of research in second language (L2) acquisition, such as learners' patterns of language use based on learner corpora (e.g., Granger, 1997, 1998; Granger, Hung & Petch-Tyson, 2002; Meunier & Granger, 2008), language education (e.g., specifying the lexical composition of pedagogical materials and corpus-informed English-medium textbooks), classroom activities, such as the use of corpus-based tools in concordancing activities (e.g., Cobb, 1997), as well as language assessment (e.g., speaking assessment by Ball & Wilson, 2002; Taylor, 2003). In vocabulary research, it has allowed for examination of lexical features and patterns in linguistic corpora, including word collocations (e.g., Partington, 1998; Sinclair, 1991), formulaic language in different registers (Simpson, 2004; Simpson & Mendis, 2003), and lexical bundles in spoken and written English (e.g., Altenberg, 1998; Biber, Conrad, & Cortes, 2004). Corpus-based research has also been

☆ The handling of the editorial process for this paper was completed by Professor Hamp-Lyons.

* Corresponding author.
E-mail addresses: sima.paribakht@uottawa.ca (T.S. Paribakht), swebb27@uwo.ca (S. Webb).

http://dx.doi.org/10.1016/j.jeap.2015.05.009
1475-1585/

Table 1
Example word families from Coxhead's (2000) Academic Word List.

Headword	Alter	Area	Assess	Assume
Family members	Alterable	Areas	Assessable	Assumed
	Alteration		Assessed	Assumes
	Alterations		Assesses	Assuming
	Altered		Assessing	Assumption
	Altering		Assessment	Assumptions
	Alternate		Assessments	
	Alternating		Reassess	
	Alters		Reassessed	
	Unalterable		Reassessing	
	Unaltered		Reassessment	
			Unassessed	

fruitful in terminology and lexicography research and practice (e.g., Collins COBUILD (Collins Birmingham University International Language Data base) English Language Dictionary (HarperCollins, 1995), and the Cambridge Advanced Learner's Dictionary, 2005).

1.2. Word frequency lists

A further fruitful application of corpus-based research has been the development of general and specialized word frequency lists, which have influenced vocabulary research, L2 vocabulary instruction, and assessment of lexical knowledge. Examples of word frequency lists include The General Service List (GSL), developed by West (1953), which comprises the 2000 most frequent English word families drawn from a written corpus of 5 million words; The University Word List (UWL) (Xue & Nation, 1984), generated for general academic purposes from four existing word lists, which comprises 836 high frequency word families, excluding GSL words; and the Academic Word List (AWL), developed by Coxhead (2000), which includes the 570 most frequent word families (Level 6 of Bauer & Nation, 1993, scale)[1] in academic written text, beyond those in the GSL. Thus, any member of a word family would be counted as 1 token in the analysis. Table 1 shows the different items that are included in the following four word families from the Academic Word List: alter, area, assess, assume.

The AWL list was drawn according to frequency and range of occurrence of words from a corpus of 3.5 million words from a variety of written academic texts in several disciplines (i.e., arts, commerce, law and science) (See Coxhead, 2011 for details.)[2], and includes words such as: indicate, approach, constitute, empirical, identify, sequence, and interpret that may be of great value for academic study, but may not be commonly known by language learners.

Each frequency list provides a certain level of lexical coverage in source texts. Lexical/text coverage for readers refers to "the percentage of running words in the text known by the readers" (Nation, 2006, p. 61). It is calculated by determining the total number of words that are known within a text, multiplying this figure by 100 and then dividing this by the number of tokens (running words) in the text (Nation, 2013). For example, if the GSL provides coverage of 80% of a text, this indicates that learners with knowledge of all 2000 GSL word families would know at least 80% of the words in that text. Research investigating coverage of the AWL in academic written text has found it to account for between 9.06% (Martínez, Beck, & Panza, 2009) and 10.00–11.60% (Chen & Ge, 2007; Cobb & Horst, 2004; Coxhead, 2000; Hyland & Tse, 2007; Vongpumivitch, Huang, & Chang, 2009; Ward, 2009) of the running words. That is, approximately 10% of words in the academic texts consisted of AWL words. Knowledge of these words should enable readers to understand the meaning of about 10% of the words in academic texts.

The AWL has been used in language education and research for over a decade and, like any other popular instrument, has inevitably been the subject of scrutiny. For example, the list has been criticized on the grounds that it is based on the GSL, which contains many high frequency academic words (Gardner & Davies, 2013), includes many words in the highest frequency lists of the British National Corpus (e.g., Cobb, 2010; Gardner & Davies, 2013; Nation, 2004), and that the words in the list have differential distribution, semantic and collocational behaviors across various fields (e.g., Cobb & Horst, 2004; Hyland & Tse, 2007).

Gardner and Davies (2013) proposed a new academic vocabulary list (AVL), comprising 3000 words. The list is based on a much larger and recent corpus (120-million-word academic subcorpus of the Corpus of Contemporary American English (COCA; Davies, 2012) from 9 academic disciplines. In compiling the list, lemmas[3] instead of word families were counted, and 'general' high frequency and technical words were removed from an academic core. The list, in both lemma and word-family versions, is available at: www.academicwords.info. The website allows users to obtain data about each AVL word (definitions,

[1] "For the creation of the AWL, a word family was defined as a stem plus all closely related affixed forms, as defined by Level 6 of Bauer and Nation (1993) scale. The Level 6 definition of affix includes all inflections and the most frequent, productive, and regular prefixes and suffixes." (Coxhead, 2000, p. 218)

[2] AWL is a smaller and more recent list than UWL but has a slightly higher coverage of academic texts. According to Coxhead (2000), the two lists have an overlap of 51% and 435 word families occur in both lists.

[3] Lemmas are "words with a common stem, related by inflection only, and coming from the same part of speech" (Gardner & Davies, 2013, p. 4).