The Journal of Language Teaching and Learning™

2016 Volume 6/Issue 2 Article 6

Comparing the Academic Word List with the Academic Vocabulary List: Analyses of Frequency and Performance of English Language Learners

K. James Hartshorn, Brigham Young University, james_hartshorn@byu.edu

Judson M. Hart, Brigham Young University, judson_hart@byu.edu

Recommended Citations:

APA

Hartshorn, K. J., & Hart, J. M. (2016). Comparing the academic word list with the academic vocabulary list: analyses of frequency and performance of English language learners. *The Journal of Language Teaching and Learning, 6*(2), 70-87.

MLA

Hartshorn, K. James, & Hart M. Judson. Comparing the academic word list with the academic vocabulary list: analyses of frequency and performance of English language learners. *The Journal of Language Teaching and Learning,* (2016). 70-87.

JLTL
www.jltl.org

The Journal of Language Teaching and Learning, 2016(2), pp. 70-87

Comparing the Academic Word List with the Academic Vocabulary List: Analyses of Frequency and Performance of English Language Learners

James Hartshorn[1], Judson Hart[2]

ARTICLE INFO

Article History:
Received January 25, 2016
Revisions completed June 24, 2016
Published June 30, 2016

Key Words:
Academic Word List (AWL)
Academic Vocabulary List (AVL)
English Language Learning
Corpus Linguistics
Implicational scaling

ABSTRACT

Although use of the Academic Word List (AWL) has been successful and extensive in English as a second language (ESL) materials development and pedagogy (Coxhead 2000, 2011), some scholars have raised concerns about possible flaws. In an effort to overcome limitations, Gardner and Davies (2014) have presented a "new Academic Vocabulary List" (AVL). While their description suggests a number of potential advantages of the AVL over the AWL, these lists have yet to undergo ecologically valid comparisons based on actual ESL learner performance. Thus, this study compares the AWL with the AVL in an effort to identify some of the most salient similarities and differences. While results suggest that the AWL and AVL do not differ significantly in their overall word frequencies or in their capacity to similarly represent the broad construct of academic vocabulary knowledge, they indicate stark differences in terms of their content and in the systematic sequencing of that content. Though much more research is needed, these findings suggest a number of potential advantages of the AVL for ESL teaching and learning.

Recently Gardner and Davies (2014) unveiled what they refer to as a “new Academic Vocabulary List” (AVL) based on analyses using the Corpus of Contemporary American English (COCA) (Davies 2012). They provided a rationale for why such a list is needed, how it was created, and how it might be used in English language education. Their description suggests possible advantages of the AVL over the Academic Word List (AWL), currently used in many ESL/EFL[3] learning contexts (see Coxhead, 2000, 2011). Despite

[1] Brigham Young University, james_hartshorn@byu.edu, 801-422-4034

[2] Brigham Young University, judson_hart@byu.edu, 801-422-4042

[3] Here after we will use ESL to refer to English as a second language (ESL) as well as to English as a foreign language (EFL).

what appear to be compelling arguments from Gardner and Davies arising from careful corpus-based analyses, these lists have yet to be compared based on actual ESL learner performance. Without actual learner data, claims regarding how various lists may be useful to students and practitioners remain more theoretical than pragmatic.

Thus, the purpose of this study is to compare the AWL with the AVL based on practical learner performance in an ecologically valid context in an effort to identify fundamental similarities and differences. In order to answer the proposed research questions, this study necessarily draws from several frequency-based analyses common to corpus linguistics (e.g., McEnery, & Hardie, 2011), best practices of second language test construction and administration (e.g., Bachman, 2010; Coombes, Davidson, O'Sullivan, & Stoynoff, 2012), and implicational scaling (e.g., Hakansson, 2013; Hatch & Lazaraton, 1991, Rickford, 2002). After a brief treatment of the most relevant literature, this paper will present our research questions, a description of our methods, our results, and a discussion of the implications of our findings.

2. Literature Review

Building on the pioneering work of scholars such as Thorndike and Lorge (1944) and West (1953), researchers developed an increased interest in creating academic word lists to facilitate language development in the 1970s. These included lists based on criteria such as word frequency and occurrence across a broad range of disciplines (e.g., Campion & Elley, 1971; Praninskas, 1972), as well as compilations of unfamiliar vocabulary appearing in student texts (Lynn, 1973; Ghadessy, 1979). Attempting to utilize the content of these lists more effectively, Xue and Nation (1984) aggregated them to form the University Word List. After fairly broad use, however, Coxhead (2000) raised concerns that the list was based on a small corpus that lacked "a wide and balanced range of topics" (p. 214) and that the processes used to select vocabulary lacked requisite systematicity.

As an alternative, Coxhead (2000) presented the Academic Word List (AWL), which has been used extensively since that time (Coxhead, 2011). In addition to its applications in the ESL classroom, the AWL has been the object of great interest for ESL materials developers and researchers. The AWL is based on frequency and range data from a corpus of 3.5 million words from academic texts. With the intent of focusing completely on academic words, rather than general vocabulary, the AWL intentionally excludes the most frequent 2000 word families identified by West (1953). The AWL includes 570 word families—defined by Coxhead (2000) as "a stem plus all closely related affixed forms" (p. 218) including inflections and derivations. She illustrates this with the word family associated with the stem concept, which includes nine additional family members such as conception, conceptually, conceptualization, and so on (p. 218).

Certainly Coxhead's work has improved the methods and criteria for the construction of academic word lists and has raised awareness of the importance of academic vocabulary development. Despite its pervasive popularity, though, the AWL has not been without its critics. Some researchers have raised concerns about the way in which word families form the foundation of the list. For example, scholars such as Nagy and Townsend (2012) have expressed apprehension that some family members may actually differ in their core meaning. For instance, in the preceding example from Coxhead (2000), the word conception may be more closely associated with becoming pregnant than a particular way of thinking about a concept, though both meanings are appropriate in their respective contexts. Other scholars have observed that academic vocabulary is not equally distributed across disciplines suggesting that the relative importance of particular words may be discipline specific and that differences in meaning may become even more salient across academic fields (e.g., Hyland & Tse, 2007; Matinez, Beck, & Panza, 2009).

Such concerns have led some researchers to abandon the quest for a comprehensive list of academic vocabulary in favor of discipline-specific lists (e.g., Hyland & Tse, 2007). Thus, many studies have been designed to produce lists for a variety of disciplines such as Business (Konstantakis, 2007), Chemistry (Valipouri & Nassaji, 2013), Computer Science (Minshall, 2013), Engineering (Gustafsson & Malstrom, 2013;

Ward, 2009), Medicine (Wang, Liang, & Ge, 2008), Nursing (Mohamad & Jin, 2013), Psychology (Yaghoubi-Notash & Janghi-Golezani, 2012), and even Applied Linguistics (Khani & Tazik, 2013). This trend represents an important and useful development in the field that is likely to benefit language educators and their students within specific disciplines.

However, we are not convinced of the advisability of assuming a mutually exclusive posture regarding vocabulary lists. We believe that a list of core academic vocabulary may still be greatly beneficial. As Gordon (2007) points out, most first-year university students in North America are undecided about their choice of what to study, including those who have already officially declared a particular major. She emphasizes that approximately 75% of these university students change their major one or more times before graduation. She claims that most students are unprepared to commit to a specific major until they have adequately explored their options and that "changing their minds is not only acceptable, but often desirable" (Gordon, p. 86). In addition to the difficulties associated with successfully steering students toward a discipline-specific vocabulary is the need for many of these same students to complete general education courses in a wide array of fields that will be associated with different vocabulary priorities. As students engage in these courses, it is quite probable that they will encounter polysemy or nuanced differences in meaning for the same word form across disciplines. Though these observations may not be as relevant outside of North America, it seems that many university-bound and first-year students may benefit from a systematic study of core academic vocabulary. In terms of materials development, even when it may be desirable to use discipline-specific lists, there may be added benefits from also drawing from a core of academic vocabulary that is applicable across all disciplines.

Because of their belief in the need for a well-designed list of core academic vocabulary, Gardner and Davies (2014) determined that the use of lemmas[4]rather than word families would make a more useful list. This was done to avoid many of the meaning-related problems associated with word families along with the expectation that knowledge of inflectional word relationships precedes knowledge of derivational word relationships (Gardner, 2007). Gardner and Davies also recognized the need to utilize a large, contemporary corpus including many academic disciplines.

In addition, Gardner and Davies (2014) considered ratio, range, dispersion, and discipline measure in the formation of the AVL. For example, they eliminated all general, high frequency words that were not at least 50% more frequent in their academic corpus of 120 million words than the non-academic portion of the corpus of 305 million words (a ratio of 1.5). They also required that the word occur "with at least 20% of the expected frequency in at least seven of the nine academic disciplines" (p. 11). Related to range, they required a dispersion of ≤ 0.80, ensuring relatively equal spread across the corpus while eliminating words that were too discipline-specific. To further decrease unwanted technical terms, they required that no word could occur "more than three times the expected frequency in any of the nine disciplines" (p. 12) that were included in their corpus of academic vocabulary.

Given this description, the most salient differences between the AWL and the AVL become more apparent and are summarized in Table 1. This comparison conveys that the AVL is based on a much larger corpus of academic texts and that it may include a broader range of academic disciplines. Beyond these observations, perhaps the most striking differences are in the AVL's use of lemmas rather than word families and in its sequencing of items from 1 to 3015 rather than grouping word family members into sublists and then ordering them alphabetically within the sublists. Other differences are also evident such as the specific methods used for excluding the highest frequency words and the most specialized vocabulary.

Despite these marked differences, one might well ask how meaningful the variations in the AWL and the AVL actually are in practice. More specifically, how likely would the use of one list versus the other

[4] Here they define a lemma as "words with a common stem, related by inflection only, and coming from the same part of speech" (p. 4).

list make a difference in English language learner performance? Some program administrators and materials developers might adopt the AVL simply based on its content, how it was constructed, or the efforts undertaken to overcome limitations apparent in previous lists. Others might assume that the two lists are similar enough that it may not matter which list is used.

Table 1
Strategic differences in the creation of the AVL and AWL

Criteria	**AWL**	**AVL**
Corpus Size	Academic corpus of 3.5 million academic words	Academic corpus of 120 million words from the COCA
Domains	28 subjects organized in to 7 general areas within 4 disciplines: Arts, Commerce, Law and Science	9 disciplines: Education, Humanities, History, Social Science, Philosophy and Religion, Law & Political Science, Science and technology Medicine and Health, Business and Finance
Selection	Word Family—a stem plus all closely related affixed forms. 570 word families with 3111 members, each of which had to occur at least 100 times in the academic corpus	Lemma—words with a common stem, related by inflection only and coming from the same part of speech. After using the exclusion techniques listed below, 3015 words were left in the list
Exclusion of high-frequency words	Exclusion of most frequent words from the General Service List	Frequency ratio of 1.5 in academic texts over non-academic texts (50% greater)
Exclusion of Technical Terms	Range: Word family members had to occur at least 10 times in each of the four disciplines and in 15 or more of the 28 subject areas included in the academic corpus	Range: Lemmas had to occur with at least 20% of expected frequency in at least 7 of the 9 disciplines; Dispersion .80; Discipline measure: a word could not occur more than 3 times the expected frequency in any of the 9 disciplines
Ordering	Based on inclusion in one of 10 sublists arranged by frequency of word family (arranged alphabetically within a sublist)	Based on frequency rank of lemmas from 1 to 3015 as sequenced within the academic corpus from the COCA

While longitudinal studies across a variety of contexts may be needed before the field can fully answer questions about the comparability of the AWL and AVL on actual L2 development, there are a number of questions that we may be able to answer now. First, some ways in which the AWL and AVL overlap or differ in their lexical constituents has yet to be reported. Such insights would further help researchers and practitioners to understand ways in which the AWL and AVL differ. Second, preliminary review of the construction of these lists suggests that two of the greatest differences between the AWL and the AVL are in how words were selected and sequenced. This raises a compelling question regarding a potential benefit for those who use the AVL resulting from its use of lemmas (rather than word families) and in its sequencing of all the lexical items in the entire list (rather than grouping them into sublists). This sequencing of the AVL words provides frequency information that seems to be diluted or lost in the AWL through the use of word families and the aggregation of words into sublists. Is it possible that by preserving frequency information throughout the list, the order of the AVL lemmas may more closely match natural patterns of acquisition of academic vocabulary and more appropriately mirror a learner's contact with such vocabulary? If so, the AVL may represent a more systematic and perhaps more appropriate sequence of lexical items for teaching and learning. Such questions deserve careful study.

Additional insight about the AWL and the AVL might be gleaned by using these lists to construct comparable instruments for eliciting language learner performance data. Such instruments could examine whether differences in the contents of these lists or the sequencing of that content would have a measurable effect on learner performance. For example, we could explore whether one list might better account for overall language proficiency. Moreover, the extent to which the sequencing of words in either list might form an implicational scale based on learner performance might reveal how well these lists mirror academic word knowledge development among ESL learners.

Implicational scaling is a statistical analysis used to identify implicational relationships among linguistic features mastered by second language learners (e.g., Hakansson, 2013). It shows which linguistic features may be the easiest (or first) to be learned as opposed to those that may be the most difficult (or last) to be learned. Hatch and Lazaraton (1991) have noted that implicational scaling can be used for examining "grammatical, lexical, and phonological features of language" and that "one of the several motivations for documenting the acquisition of these features relates to language teaching" (p. 204). They further explain that those who design syllabi and L2 language learning materials use implicational scaling to document learning sequences as demonstrated by actual learner language in order to prevent the need to rely on a fallible sense of intuition regarding the best way to sequence teaching.

Thus, scholars have used implicational scaling to study a variety of language phenomena to identify accuracy orders that might facilitate L2 teaching and learning. These include examining accuracy orders associated with linguistic features ranging from phonological development, (e.g., Nagy, Moisset, & Sankoff, 1996; Trofimovich, Gatbonton, & Segalowitz, 2007), morphosyntactic structures (e.g., Algady, 2013; Pienemann, 1998; Pienemann & Mackey, 1993), listening comprehension strategies (Young, 1997) as well as vocabulary development based on frequency levels (e.g., Ozturk, 2015; Read, 1988; Schmitt, Schmitt & Clapham, 2001). With specific reference to vocabulary development, Ozturk (2015) has observed "Word frequency has long been a major guiding principle in setting lexical targets for
L2 learners and it is assumed that learners should and will proceed according to frequency.

The preceding review of literature has suggested the value of an academic vocabulary list for pedagogy and materials development and has identified possible benefits of the new AVL. It describes the need to compare the AWL and AVL in terms of their lexical constituents and relative frequencies and suggests the benefits of using learner data to better understand how knowledge of these lists might account for language proficiency and whether performance levels may be scalable. With these considerations in mind, we formed the following research questions.

1. How do the vocabulary in the AVL and AWL differ in terms of their lexical constituents and their relative frequencies?
2. How well does knowledge of academic vocabulary, as it appears in the AVL and the AWL, account for language proficiency? Do the AVL and the AWL account for language proficiency equally well?
3. How systematic is the ordering of the AVL and AWL? Can knowledge of high, moderate, and lower frequency academic vocabulary as found in the AVL and AWL form an implicational scale? Are the AVL and the AWL equally scalable?

3. Methodology

This section describes processes and procedures used to elicit data. This includes a brief description of the lexical analyses designed to answer the first research question. It also explains the formation of an AWL test and an AVL test and associated elicitation procedures used to answer the second and third research questions.

3.1. A Lexical Comparison of the AVL and AWL

In order to answer the first research question, we conducted a basic lexical analysis of the AWL and AVL to determine the extent to which the two lists overlap and differ. We also examined frequency information for each word in each list. We used Vocabprofile online and the COCA to conduct our analyses (Cobb, 2014; Heatley, Nation, & Coxhead, 2002).

3.2. The AVL and AWL Test Construction

In order to answer the second and third research questions, we constructed two similar tests of academic vocabulary, one based on the AWL and one based on the AVL. This section will provide a description of how these tests were constructed including the choice of target words, the selection of distractors, and how the test item stems were created.

3.2.1. Subdividing the Lists

While our intent was to follow the same processes in creating the AWL and AVL tests, we recognized constraints resulting from differences between the lists. We also recognized issues that would affect how the tests could be administered. For example, we concluded that for practical reasons we would only be able to test a sample of each list. For selection of words from the AWL, we first divided the AWL based on the same curricular division used in the intensive English program (IEP) where this study was conducted. This meant Sublists 1-9 were divided in half. This division was at the boundary between the word family 30 and 31 as they appeared in the list alphabetically. Since Sublist 10 only includes 30 words, it was not divided. This created a total of 19 bands of 30 word families from the AWL.

With the goal of making the AWL and AVL tests as similar as possible while reflecting important differences in the respective lists, we determined that the AVL test would also include 19 bands as did the AWL test. However, a preliminary inspection of the AVL suggested two important differences between lists. First, the AWL seemed to have more words per family than the AVL. Second, the vocabulary toward the end of the AVL seemed much less frequent than the words at the end of the AWL. This was verified using COCA data to compare the mean word frequency (measured in millions) from the 95 word family members included in Sublist 10 of the AWL (M=4,130.600, SD=4,915.257) and the final 95 words from the AVL (M=927.758, SD=2,273.289), t(188)= 5.764, p<.001.

Therefore, in a step toward making the frequencies represented by the two tests more similar and to adjust for larger word families in the AWL, we tentatively set the width of the AVL bands at 100 words. While this only represented 1900 lemmas (63%) of the 3015 words in the AVL, this seemed to be an appropriate way to make the tests somewhat more balanced. Using this breakdown, we also divided the two lists into the high, moderate, and lower frequency groups needed to answer the third research question. Because we sought to optimize the balance between maximizing the size of our three frequency groupings and the distance between one group to another, we determined to create gaps between the high, moderate, and low frequency groups. Thus, for both lists, we defined the highest frequency group as bands 1-4, the moderate frequency group as bands 9-12, and the lowest frequency group as bands 17-19. To help determine the appropriateness of our tentative decision to divide the AVL into 19 bands of 100 words, we compared the ends of the two tests (i.e., the low frequency words for each test made of the words in Bands 17, 18, 19). Analysis revealed that the difference between mean frequency at the end of the AVL test (M=2,550, SD=4,356) and the end of the AWL test (M=4,777, SD=5,818), was not statistically significant, t(58)=-1.679, p=.099. This provided some additional support for the previous decision to limit our testing to the first 19 bands of 100 words in the AVL.

3.2.2. Selecting Words for the AWL Test

The 30 word families within each band of the AWL were expanded to include all word-family members (3111 words). Since no separate frequency information is provided for words within a given AWL sublist, these expanded sets from each sublist were placed in a randomizer. The first ten words randomly selected were included in our test to represent that band. Words selected from the same word family were used so long as the words were lexically or grammatically distinct. The first ten words randomly selected for Band 1 (Sublist 1A) are displayed in Table 2. Note that beneficial and financier happened to be included along with benefit and finance despite belonging to a family already represented since they met the randomized conditions for selection previously described. Some words selected from the AWL also occur in the AVL. This is also illustrated in Table 2, which displays the COCA rank[5] and the AVL rank for words that overlap.

Table 2
AWL words selected for Sublist 1 (Band 1)

Word	**COCA Rank**	**AVL Rank**
consistency	394	816
benefit	2247	-
finance	2825	-
constitution	4504	-
beneficial	5042	838
export	5361	1011
environmentally	6851	-
financier	13555	-
derivation	21650	2479
uneconomical	41933	-

3.2.3. Selecting Words for the AVL Test

Because pilot testing with some of the students included the first ten words from the AVL bands, we used words 11-20 from each band of 100 words from the AVL to create the instrument used in this study. The first ten words selected are included in Table 3, which also indicates the AVL rank, the COCA rank, and whether or not the word also appears in the GSL or AWL.

[5] Both the COCA rank and the AVL rank are ordered from most frequent to least frequent within each list.

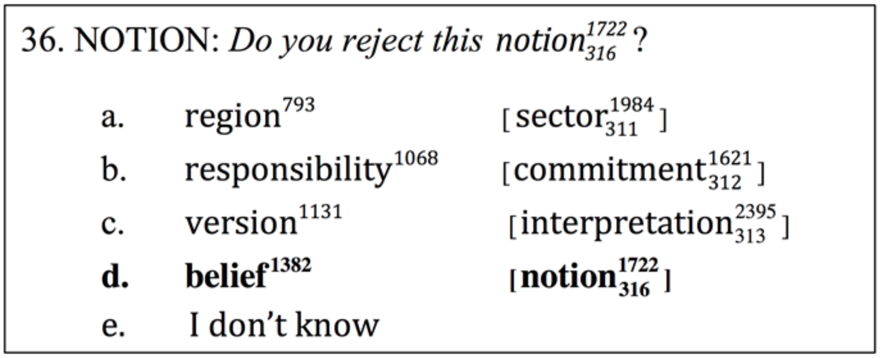
Table 3
First 10 Words selection for band 1 of the AVL test

AVL	Word	COCA Rank	Other Lists
11	Important	266	GSL
12	process	391	AWL
13	use	428	GSL
14	development	448	GSL
15	data	559	AWL
16	information	314	GSL
17	effect	427	GSL
18	change	356	GSL
19	table	539	GSL
20	policy	388	AWL

3.2.4. The Construction of Test Items

In order to address the research questions, we needed to test a large volume of vocabulary. However, since each test consisted of 190 items, making the test items concise was a high priority in order to minimize testing fatigue. As a result, academic word knowledge was operationalized as the ability to substitute an academic word with a higher frequency word or phrase that matched the meaning of the targeted word. Items on both tests were similarly constructed following the procedure outlined below. This result is illustrated in Figure 1 with the word "notion" from the fourth band of items on the AVL test form. The source distractors and targeted words are presented in brackets. Superscripts denote COCA rank and subscripts denote the AVL rank.

a. The targeted word is presented uninflected and in capital letters.
b. The word is shown used in a minimally supported stem. Minimally supported means that the grammatical constraints of the word are clear according to its use but there is minimal context support for its meaning.
c. Researchers selected the word used as the correct answer by finding the targeted word in the online COCA corpus tool. The COCA rank for the targeted word was noted. This corpus tool also identifies words that share meaning with the target word and these words' COCA rank. The correct answer was selected from those words that shared meaning with a comparable or lower rank than the targeted word.
d. Selection of distractors was similarly proceduralized. First for the AVL test, words that were adjacent to the targeted word that also shared part of speech with the targeted word were identified. For the AWL test, distractors were selected from word families from the same sublist as the targeted word. The distractors selected also matched the targeted word's part of speech.
e. Distractors were also downgraded as described above. Ultimately distractors were selected to have a COCA ranking similar to the ranking of the correct answer.
f. Test takers were presented with the answer choice of 'e. I don't know' so as to minimize detriment to the spectrum of lower proficiency language learners which took the test.



36. NOTION: *Do you reject this notion*$^{1722}_{316}$?

a. region^{793} [sector$^{1984}_{311}$]

b. responsibility^{1068} [commitment$^{1621}_{312}$]

c. version^{1131} [interpretation$^{2395}_{313}$]

d. belief^{1382} [notion$^{1722}_{316}$]

e. I don't know

Figure 1. Sample item with COCA ranking (superscript) and AVL ranking (subscript)

3.3. Participants

Examinees included students enrolled in an IEP prior to the beginning of a new semester. Proficiency levels ranged from novice-high to advanced-high on the ACTFL scale. First language backgrounds included Spanish (48%), Korean (14%), Portuguese (13%), Chinese (11%), Japanese (7%), Russian (3%), Arabic (1%), French (1%), Italian (1%), and Mongolian (1%). Ethics approval for this study was received from our university internal review board.

3.4. Data Elicitation

Each examinee took both forms of the test in very similar circumstances. To minimize testing fatigue, students were given the two tests on different occasions with a few days between occasions. Nevertheless, students were randomly assigned which test to take first in order to minimize any possible effect of test order. In both cases, the tests were given in conjunction with other testing, but the vocabulary test was the first test given on both occasions. Both administrations of the tests were tightly proctored. Examinees had 80 minutes to complete the 190 items on both test forms respectively. Students used computer-scanned answer sheets to record their responses which helped ensure accurate scoring.

3.5. Test Performance

Since our aim in constructing the AWL and AVL tests was to use the same processes without muting the inherent differences between the lists, some functional differences between the tests were anticipated. Nevertheless, before attempting to answer our research questions, we will briefly examine the relative functioning of the two tests. A group of students (N=205) ranging from novice to advanced-high proficiency took both the AVL test and the AWL test during a regularly scheduled exam period. One point was awarded for each of the 380 items answered correctly on both tests. The correlation between the two distributions of raw test scores was .913 (p<.001) with an AWL test mean of 104.961 (SD=36.784) and an AVL test mean of 114.110 (SD=40.375). This suggests that despite substantial differences between the lists, the two tests functioned similarly in assessing the construct of academic vocabulary knowledge.

3.6. Implicational Scaling

The AWL and AVL tests described above were also used to address the third research question regarding whether student knowledge of the academic vocabulary based on the AWL and AVL tests could form an implicational scale. For this study, the scalability of high, mid, and lower[6] frequency groupings of academic vocabulary as presented in the AVL and AWL was examined. Our assumption was that student performance based on both tests would be scalable according to the following: mastery of low frequency words would predict knowledge of mid frequency words, and mastery of mid frequency words would predict knowledge of high frequency words, depicted as low frequency words ⊃ mid frequency words ⊃ high frequency words.

3.6.1. Mastery Criteria for Implicational Scaling

In order to understand analyses associated with implicational scaling, familiarity with several concepts is necessary. The first is establishing an appropriate level of accuracy as a definition of mastery. If the level of accuracy is set too high, attempts to detect scalability may be undermined by limitations in elicitation instruments or procedures. If the standard is set too low, the results would no longer reflect mastery—they would be fraught with error and become much more difficult to interpret. Ellis (1994) indicates that most studies in second language acquisition have used a criterion of 80 to 90%. This margin of 10 to 20% allows for some limitations in elicitation instruments or procedures without having them adversely affect the outcome of the analysis. Based on this recommendation from Ellis and other researchers, the accuracy level was set at 90%. In practical terms, this would mean that the student would need to respond correctly on 90% of the items targeting words at a particular frequency level before we would assume student mastery of the vocabulary at that level.

3.6.2. Analyses for Implicational Scaling

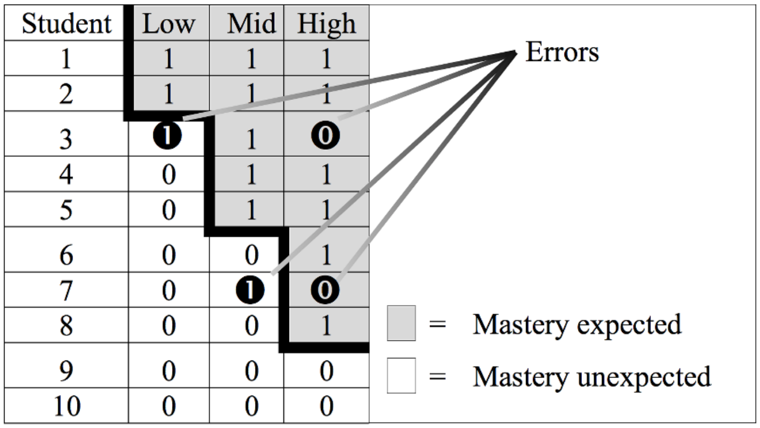
Following guidelines provided by Hatch and Lazaraton (1991), the coefficient of reproducibility (Crep), the minimum marginal reproducibility (MMrep), the % of improvement, and the coefficient of scalability (Cscal) were calculated. Each of these is described briefly. Since implicational scaling is one of a few statistical procedures that many professionals in the field feel least able to interpret (Loewen, Lavolette, Spino, Papi, Schmidtke, Sterling, & Wolff, 2014), we will attempt to be as explicit and transparent as possible.

The Crep shows how well learner performance can be predicted by student rankings. This coefficient must be greater than .90 for the scale to be valid and is calculated as:

$$C_{rep} = 1 - \frac{\text{number of errors}}{\text{(number of subjects)(number of items)}}$$

Here the number of subjects simply refers to the total number of participating students. There are three items (i.e., high, mid, and lower frequency vocabulary). The number of errors refers to those cases where accuracy orders do not fit the model being tested. Consider the following example in Figure 2, which illustrates the hypothetical performance of 10 students on three frequency levels of academic vocabulary: low, mid, and high. If 1 represents mastery, Students 1 and 2 have demonstrated mastery at all three levels but that Students 9 and 10 achieved none of the levels. Expected mastery is based on the number of levels achieved, so for Student 1, it would be all 3 levels, for Student 3 it would be 2 levels, and for Student 8, it would be 1 level. Each deviation from what is expected constitutes an error. Thus, we see two errors from Student 3 and two errors from Student 7.

[6] All words in this study could be considered *high frequency*. To differentiate groupings, the term *lower frequency* is used but is relative.



Student	Low	Mid	High
1	1	1	1
2	1	1	1
3	❶	1	⓿
4	0	1	1
5	0	1	1
6	0	0	1
7	0	❶	⓿
8	0	0	1
9	0	0	0
10	0	0	0

Errors
▨ = Mastery expected
□ = Mastery unexpected

Figure 2. Sample matrix for implicational scaling

The MMrep is a measure of how well learner performance can be predicted without considering errors. The maximum marginals are calculated by adding all of the 1s and 0s within each column to determine which is greater. The larger of the two numbers from each column is summed. With this information, the MMrep can be calculated as:

$$MM_{rep} = \frac{\text{maximum marginal}}{\text{(number of subjects)(number of items)}}$$

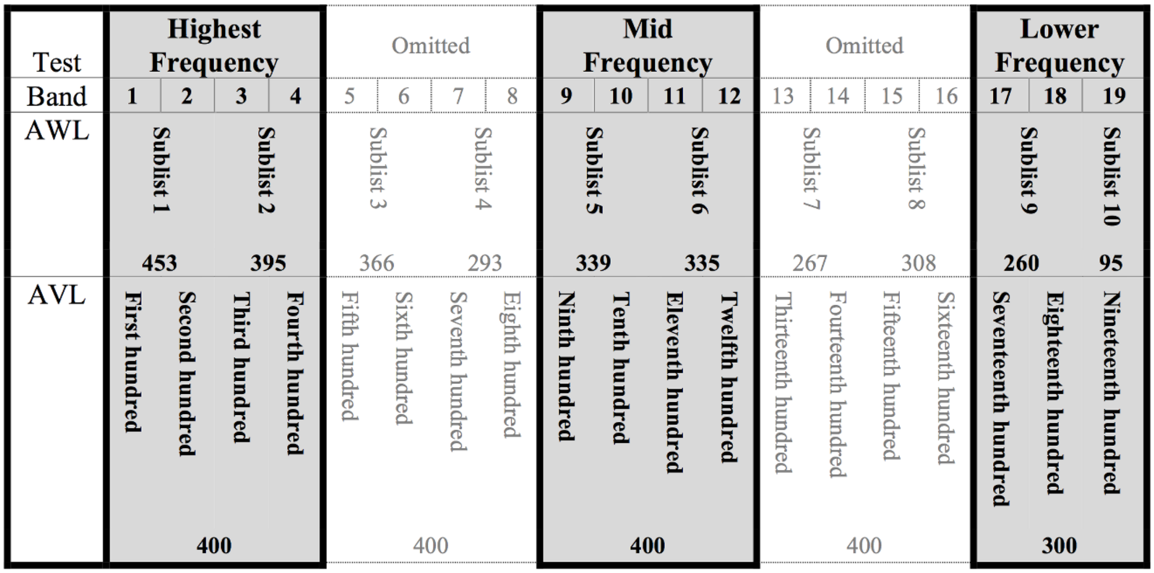
The percent of improvement in reproducibility simply measures the difference between the Crep and the MMrep, calculated as Crep - MMrep. Finally, the Cscal shows whether the items (i.e., high, mid, or low frequencies levels of academic vocabulary) can actually form an implicational scale and is calculated as:

$$C_{scal} = \frac{\text{\% improvement in reproducibility}}{1\text{-} MM_{rep}}$$

The Cscal must be above .60 before researchers can claim scalability. Thus, learner performance on the three frequency levels of academic vocabulary could be considered valid and scalable only if the following are true: Crep >. 90 and the Cscal > .60.

3.6.3. Data Used for Implicational Scaling

In order to answer the third research question, the AWL and AVL tests were divided into sections representing high, mid, and lower frequency vocabulary as illustrated in Figure 3. For each test, the highest frequency academic vocabulary was defined as those 40 words from Bands 1-4 of the AWL or the AVL respectively. To ensure separation of frequency levels, Bands 5-8 were omitted, and the mid frequency group was defined as those words in Bands 9-12. Similarly, Bands 13-16 were omitted, and the lower frequency group was defined as those words from Bands 17-19. This represented an attempt to balance the need to include as many words as possible to maximize reliability without extending the group beyond what would be appropriate for its frequency grouping.



Test	Highest Frequency				Omitted				Mid Frequency				Omitted				Lower Frequency		
Band	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
AWL	Sublist 1 453		Sublist 2 395		Sublist 3 366		Sublist 4 293		Sublist 5 339		Sublist 6 335		Sublist 7 267		Sublist 8 308		Sublist 9 260		Sublist 10 95
AVL	First hundred	Second hundred	Third hundred	Fourth hundred	Fifth hundred	Sixth hundred	Seventh hundred	Eighth hundred	Ninth hundred	Tenth hundred	Eleventh hundred	Twelfth hundred	Thirteenth hundred	Fourteenth hundred	Fifteenth hundred	Sixteenth hundred	Seventeenth hundred	Eighteenth hundred	Nineteenth hundred
	400				400				400				400				300		

Figure 3. Words selected for implicational scaling

4. Results

Various analyses were utilized in an effort to answer the three research questions. These questions addressed potential differences between the AVL and AWL in terms of their basic constituents, their ability to predict proficiency, and in their scalability. Each is presented below.

4.1. The AVL and AWL Constituents and Their Frequencies

The first research question addresses how the AWL and AVL differ in terms of their basic constituents. Surprisingly, only 31.05% of the AVL overlaps with the AWL. Approximately 26.87% of the AVL is made up of the first 2000 most frequent words as found in the GSL (West, 1953). These include various words such as distinguish, systematic, and efficiency. In addition, 42.07% of the AVL is not included in the GSL or in the AWL. This includes words such as verify, optimal, and configuration. The observation that more than two thirds of the contents of the AWL and AVL are mutually exclusive underscores a fundamental difference between these lists.

Moreover, since the AWL includes 570 headwords, which correspond to a total of 3111 different word family members, we find 5.46 words per family in the AWL. On the other hand, the AVL includes 3015 words or lemmas ordered from 1 to 3015 based on frequency. These can be organized into 1710 word families, resulting in 1.76 words per family. Thus, while the total number of words in each list is similar, the AVL has approximately three times the word families as the AWL.

In order to further understand differences between these lists, we also compared the word frequencies from each of the sublists of the AWL (expanded to include all word family members within each sublist) with an equivalent number of AVL words as were included in each of the AWL sublists (because the lists number have different totals, no comparison could be made for sublist 10). An ANOVA comparing the AWL and AVL revealed a significant difference across sublists, F(8,5997)=168.248, p=<.001, ηp2= .183. These differences are illustrated in Figure 4, which displays mean frequencies (based on occurrences per million words from the COCA). While these differences across sublists were anticipated, we were surprised to see the accompanying analysis of overall word frequency comparing AWL (M=6,899, SD=13,838) and AVL (M=8,830, SD=18,936) show no significant difference, F(1,5997)=1.502, p=.220. This suggests that, in terms of overall word frequency, list differences are actually negligible. Thus, frequency differences between sublists can only be accounted for in how the words in the respective lists are sequenced.

4.2. Using the AVL and AWL to Explain Proficiency

To answer the specific research question regarding the potential of these lists to account for proficiency levels, raw scores from the respective AWL and AVL vocabulary tests were used independently as the explanatory variable, and scores from in-house placement tests of reading, writing, listening, speaking and grammar for 8 proficiency levels (0-7) were used as the response variable for a simple linear regression. The analysis using the AVL produced an adjusted R2 of .557 (p<.001) and the analysis using the AVL produced an adjusted R2 of .563 (p<.001). These results suggest that both tests performed comparably and that each successfully accounted for more than half of the variability associated with student proficiency.

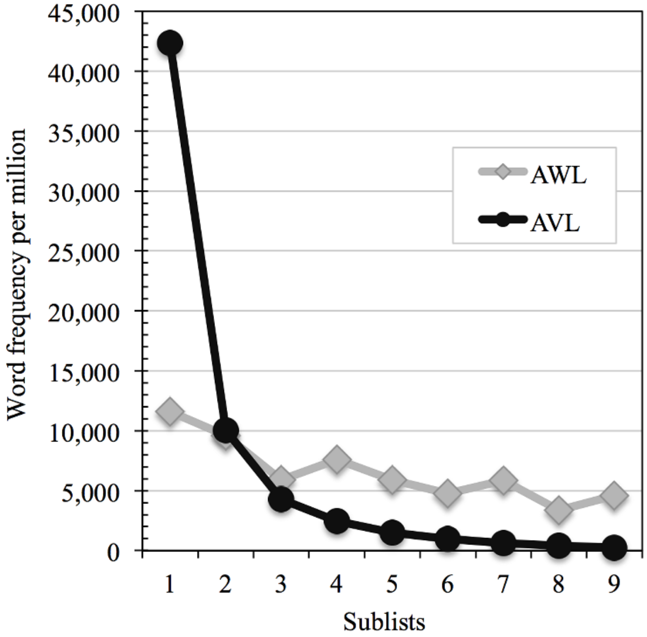


Figure 4. Mean word frequencies plotted by sublist

4.3. Implicational Scaling

4.3.1. Scaling of the AWL Test Results

Using a 90% mastery criterion, performance data from the 218 learners on the AWL test did not produce an implicational scale according to the following[7] calculations:

$$C_{rep} = 1 - \frac{\text{26 Errors}}{\text{(218 subjects)(3 items)}} = .960$$

$$MM_{rep} = \frac{209+210+218}{\text{(218 subjects)(3 items)}} = .974$$

[7] Although values included in these and subsequent formulae appear as rounded numbers, more precise values were used for actual calculations and differ slightly from these values used only for illustration.

$$C_{scal} = \frac{-.014}{1-.974} = -.529$$

Of the 13 learners who exhibited mastery of the high frequency words, the mid frequency words, or both, none were without errors in the scale's accuracy order. Moreover, none of the students demonstrated mastery of the lower frequency vocabulary at the 90% mastery level.

4.3.2. Scaling of the AVL Test Results

Unlike the AWL, performance of the 223 learners with a 90% mastery criterion on the AVL test produced an accuracy order that could be considered both valid and scalable according to the following:

$$C_{rep} = 1 - \frac{\text{16 Errors}}{\text{(223 subjects)(3 items)}} = .976$$

$$MM_{rep} = \frac{216+197+188}{\text{(223 subjects)(3 items)}} = .898$$

$$C_{scal} = \frac{.078}{1-.898} = .765$$

As expected, mastery of the lower frequency vocabulary suggested mastery of the mid frequency vocabulary, and mastery of the mid frequency vocabulary suggested mastery of the highest frequency vocabulary (i.e., low frequency ⊃ mid frequency ⊃ high frequency).

4.3.3. A Posteriori Analyses

Though the research question specifically targeted just the 90% mastery level, we were concerned that the 90% mastery criterion may have been too stringent amid possible limitations in the elicitation instrument or procedures. Therefore, additional analyses were calculated to determine whether a different mastery criterion would result in scalable results for the AWL test. This included 80%, 70%, and 60% mastery levels respectively. The results of these and the previous analyses are summarized in Table 4. The results of each level of mastery for the AVL test were scalable and produce the same expected accuracy order.

Table 4
Test of implicational scaling by percent of mastery

	Academic Vocabulary List			**Academic Word List**		
%	C_{rep}	MM_{rep}	C_{scal}	C_{rep}	MM_{rep}	C_{scal}
90	.976	.898	*.765	.960	.974	-.529
80	.970	.771	*.870	.921	.887	.297
70	.973	.648	*.924	.924	.722	†.725
60	.961	.632	*.894	.939	.587	*.852

*Scalable in expected order † Scalable but not in expected order

However, the first sign of scalability for the results of the AWL test appear at the 70% mastery criterion. Nevertheless, the scale ordering was unanticipated and violated our expectation. Knowledge of the low frequency vocabulary suggested mastery of the high frequency vocabulary, and knowledge of the high frequency vocabulary suggested mastery of the moderate frequency vocabulary (i.e., low frequency

⊃ high frequency ⊃ mid frequency). However, at the 60% mastery level, we finally observed scalable results from the AWL test data that fit our expectation. This produced an accuracy order of low frequency ⊃ mid frequency ⊃ high frequency.

5. Discussion

5.1. Similarities and Differences

This study examined three research questions designed to reveal possible differences between the AWL and the AVL. These findings should be of great interest to ESL/EFL materials developers, practitioners, program administrators, and researchers. Both lists appear equally useful in measuring the broad construct of academic vocabulary knowledge, and the fact that both tests independently accounted for more than half of the variability associated with proficiency is impressive. These findings suggest important similarities between the AWL and the AVL that are visible at a macro level.

However, closer examination reveals more differences than similarities. Though results showed that there is some overlap between the two lists, 68.95% of the AWL and AVL is mutually exclusive. This difference was greater than anticipated. This is partially due to the fact that just over a quarter of the AVL words also appear in the GSL. Since the AWL was built on top of the GSL with the expectation that learners could transition from the GSL to the AWL, none of the GSL words appear in the AWL by design. Nevertheless, the rationale for excluding the most frequent words from a corpus from the early 1900s may no longer be justifiable since analyses of the AWL have shown that it includes many of the words in contemporary high-frequency lists (e.g., Cobb 2010; Neufeld *et al*. 2011; Schmitt & Schmitt 2012).

Beyond the GSL, however, the AVL contains many additional words that are not part of the AWL. In terms of list structure, the AVL begins at a much higher frequency than the AWL but then ends at a lower frequency. While there is no statistically significant difference between the two lists in terms of their overall frequencies, there is a fairly dramatic difference in how frequency distributions are sequenced. In this regard, the AVL appears to be more systematic from one section of the list to another while the AWL seems less systematic in terms of incremental changes in frequency across sublists. The most obvious possibilities for this observation include the arrangement of the list around word families and the aggregation of word families into sublists. These procedures in the creation of the AWL may have diluted frequency differences from one sublist to the next. If frequency is a desirable feature for sequencing the teaching and learning of vocabulary, then the processes that underlie the development of the AVL may provide some advantages.

Another important observation is that the AVL includes three times the number of word families or headwords (1710) compared to the AWL (570). This appears to give the AVL less depth but much more breadth compared to the AWL. Unlike the AWL, which is based on form alone, this additional breadth provided by the AVL is presented in lemmas, which includes part of speech. While this choice to use lemmas will not eliminate all polysemy, it adds a great deal of clarity for practitioners and material developers. For example, within the first 10 words of the AVL we encounter words like *study*, *research*, *level*, and *result*. Knowing that each of these are nouns in this segment of the AVL, rather than verbs, makes the task for the practitioner or materials developer much more focused. While *study* and *level* never appear in the AVL as verbs, the other words emerge much later in the list as verbs (i.e., *result*, 189; *research*, 784). The extent to which this use of lemmas might facilitate academic vocabulary development is worth additional study.

5.2. Limitations and Future Research

Because of inherent differences between the AWL and the AVL some of the rational decisions made in the design of this study could be viewed both as strengths as well as limitations. For example, in our attempt to equalize frequencies across lists, account for large word families in the AWL, and to balance the testing experience of those taking the two tests, we limited the AVL test to the first 19 bands of 100 words. While this was a reasonable decision, it did not provide a comparison of the two complete lists; this is something worth pursuing in additional study. Moreover, because of practical constraints, we operationalized word knowledge in fairly simplistic terms. Further study could examine multiple and deeper aspects of word knowledge.

In addition, we recognized that it is conceivable that some of the differences between the AWL and AVL observed in this study may be attributed to the different corpora that were used in the construction of these lists and in our analyses of them. For example, the AWL was developed with an academic corpus of 3.5 million words (Coxhead, 2000), while the AVL was created with a more recent academic corpus of 120 million words. It may be possible that the size or currency of these corpora may have impacted our results. If this is true, however, it raises concerns about the use of corpora that may be older or smaller.

While this study examined a number of general ways in which the AWL and the AVL differ structurally, very little in this study addresses the actual content of these lists. Since the contents of the AWL and AVL are more different than similar, additional study needs to provide a more detailed analysis of how the contents of these lists differ and how those differences may affect language development in various teaching and learning contexts. Finally, we need to understand the ways in which a focus on lemmas rather than word forms may help language learners in their vocabulary development.

6. Conclusion

This study explored a number of similarities and differences between the AWL and the AVL that should be of interest to a variety of researchers as well as to ESL materials developers, practitioners, and program administrators. At the macro level, the AWL and AVL do not differ significantly in their overall word frequencies; nor do they differ in their capacity to account for differences across learner proficiency levels. However, this study also shows a number of ways in which the AWL and AVL differ. In comparison, the AVL seem much more systematically sequenced. This is evident both in in terms of its construction as well as through data elicited from ESL learner performance. Much more study is needed to further identify important differences between the AWL and AVL and help us to understand how these differences may affect academic vocabulary development in various teaching and learning contexts.

References

Algady, D. (2013). *The acquisition of relative clauses: How to second language learners of Arabic do it?* An unpublished dissertation, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography, 6*, 253–279.

Campion, M. & Elley, W. (1971). *An Academic Word List*. Wellington: New Zealand Council for Educational Research.

Cobb, T. (2015). Web Vocabprofile. (version 2) [software]. Available from http://www.lextutor.ca/vp/ [an adaptation of Heatley, Nation & Coxhead (2002)].

Coombes, C., Davidson, P., O'Sullivan, B., & Stoynoff, S. (Eds.). (2012). The Cambridge guide to second language assessment. Cambridge: Cambridge University.

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

Coxhead, A. (2011). The Academic Word List 10 years on: research and teaching implications. *TESOL Quarterly, 45*, 355–62.

Davies, M. (2012). Corpus of Contemporary American English (1990–2012). Retrieved from http://corpus.byu.edu/coca/.

Ellis, R. (2008). *The study of second language acquisition*. Oxford: Oxford University Press.

Gardner, D. & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*, 305-327.

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: a critical survey. *Applied Linguistics, 28*, 241–65.

Ghadessy, P. (1979). Frequency counts, word lists, and materials preparation: a new approach. *English Teaching Forum, 17*, 24–7.

Gordon, V. N. (2007). *The undecided college student: An academic and career advising challenge* (3nd. ed.). Springfield, IL: Charles C. Thomas.

Hakansson, G. (2013). Implicational Scaling In P. Robinson (Ed), *Routledge Encyclopedia of Second Language Acquisition* (pp. 293-294). New York: Routledge,

Hatch, E. & Lazaraton, A. (1991). *The research manual: design and statistics for applied linguistics*. Rowley, MA: Newbury House.

Heatley, A., Nation, I.S.P. & Coxhead, A. (2002). Range and Frequency programs [software]. Available at http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

Hyland, K. & Tse, P. (2007). Is there an ''Academic Vocabulary''?, *TESOL Quarterly, 41*, 235–53.

Konstantakis, N. (2007). Creating a business word list for teaching English, *Estudios de lingüística inglesa aplicada, 7*, 79-102.

Lightbown, P. M., & Spada, N. (1999). *How languages are learned*. Oford: Oxford University Press.

Loewen, S., Lavolette, E., Spino, L.E., Papi, M., Schmidtke, J., Sterling, S., & Wolff, D. (2014). Statistical literacy among applied linguists and second language acquisition researchers. *TESOL Quarterly, 48*, 360-388.

Lynn, R. W. (1973). Preparing word lists: a suggested method. *RELC Journal, 4*, 25–32.

Martinez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes, 28*, 183–198.

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Minshall, D. E. (2013). *A Computer Science Word List*. Unpublished MA thesis submitted to Swansea University, Singleton Park, Wales.

Mohamad, A. F. N. & Jin, N. Y. (2013). Corpus-based studies on nursing textbooks. *Advances in Language and Literacy Studies, 4*, 21-47.

Nagy, N., Moisset, C., & Sankoff, G. (1996). On the Acquisition of Variable Phonology in L2. *University of Pennsylvania Working Papers in Linguistics, 3*, 111-126.

Nagy, W. & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly, 47*, 91–108.

Ozturk, M. (2015) Vocabulary growth of the advanced EFL learner. *The Language Learning Journal, 43*, 94-109.

Pienemann, M. (1998). *Language Processing and Second Language Development: Processability Theory*. Amsterdam: John Benjamins.

Pienemann, M., & Mackey, A. (1993). An empirical study of children's ESL development and rapid profile. In P.

McKay (Ed.), ESL development: Language and literacy in schools (pp. 115-259). Canberra: Commonwealth of Australia and National Languages and Literacy Institute of Australia.

Praninskas, J. (1972). *American University Word List*. London: Longman.

Prince, P. (2012). Toward an instructional programme for L2 vocabulary: Can a story help?' *Language Learning & Technology, 16*, 103-120.

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal, 19*, 12–25.

Rickford, J.R. (2002). Implicational scales. In J.K. Chambers, P.J. Trudgill and N. Schilling-Estes (Eds.), *The handbook of language variation and change* (pp. 142-167). Oxford: Blackwell.

Schmitt, N., Schmitt, D, & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18,* 55–88.

Thorndike, E. L. & Lorge, I. (1944). *The teacher's word book of 30000 words*. New York: Columbia Teachers College.

Trofimovich, P., Gatbonton, E., & Segalowitz, N. (2007). A dynamic look at L2 phonological learning: Seeking processing explanations for implicational phenomena. *Studies in Second Language Acquisition, 29,* 407-448.

Wang, J., Liang, S., & Ge, G. (2008). Establishment of a medical academic word list. *English for Specific Purposes, 27*, 442–458.

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

Xue, G. & Nation, I. S. P. (1984). A university word list,' *Language Learning and Communication, 3*, 215–29.

Yaghoubi-Notash, M. & Janghi-Golezani, M. (2013). From frequency to instructional order: insights from narrow-angle corpus of psychology RA introductions. *Theory and Practice in Language Studies, 3*, 1034-1039.

Young, M. Y. C. (1997). A serial ordering of listening comprehension strategies used by advanced ESL learners in Hong Kong. *Asian Journal of English Language Teaching, 7*, 35-53.