Applied Linguistics Advance Access published August 2, 2013

Applied Linguistics 2013: 1–24
doi:10.1093/applin/amt015

A New Academic Vocabulary List

*DEE GARDNER and MARK DAVIES

Department of Linguistics and English Language, Brigham Young University
*E-mail: dee_gardner@byu.edu

This article presents our new Academic Vocabulary List (AVL), derived from a 120-million-word academic subcorpus of the 425-million-word Corpus of Contemporary American English (COCA; Davies 2012). We first explore reasons why a new academic core list is warranted, and why such a list is still needed in English language education. We also provide a detailed description of the large academic corpus from which the AVL was derived, as well as the robust frequency and dispersion statistics used to identify the AVL. Our concluding case studies show that the AVL discriminates between academic and other materials, and that it covers ~14% of academic materials in both COCA (120 million+ words) and the British National Corpus (33 million+ words). The article concludes with a discussion of how the AVL can be used in settings where academic English is the focus of instruction. In this discussion, we introduce a new web-based interface that can be used to learn AVL words, and to identify and interact with AVL words in any text entered in the search window.

1. ACADEMIC VOCABULARY KNOWLEDGE

Academic vocabulary knowledge is recognized as an indispensable component of academic reading abilities (Vacca and Vacca 1996; Corson 1997; Biemiller 1999; Nagy and Townsend 2012), which, in turn, have been directly linked to academic success, economic opportunity, and societal well-being (Goldenberg 2008; Ippolito *et al*. 2008; Jacobs 2008). This central role of academic vocabulary in school success is true for both native and non-native speakers of English, and at all grade levels, including primary (Chall 1996; Biemiller 2010), middle-school (Townsend and Collins 2009), secondary (Vacca and Vacca 1996), and higher education (Schmitt *et al*. 2011). In fact, control of academic vocabulary, or the lack thereof, may be the single most important discriminator in the 'gate-keeping' tests of education: LNAT, UKCAT, HAT, BMAT, ELAT (in Great Britain), SAT, ACT, GMAT, LSAT, GRE, MCAT (in the USA and Canada), STAT, UMAT, GAT, GAMSAT (in Australia and New Zealand), TOEFL, Michigan (for non-native English speakers), and many others. Insufficient academic vocabulary knowledge has also been strongly associated with the oft-cited 'gap' in academic achievement that exists between certain groups of students—primarily the economically disadvantaged and English language learners—and their grade-level peers (Hart and Risley 1995; Chall 1996; Hiebert and Lubliner 2008; Neuman 2008; Biemiller 2010; Lesaux *et al*. 2010; Townsend *et al*. 2012).

Acknowledgement by educators and language experts of such high-stakes academic needs in primary, secondary, and higher education has led to a proliferation of books and articles about the vocabulary of schooling—what its characteristics are, why it is a problem for learners, how it should be taught, and so forth (e.g. Beck *et al.* 2002; Graves 2006; Nation 2008; Zimmerman 2009; Bauman and Graves 2010; Biemiller 2010; Carter 2012; Nagy and Townsend 2012; Gardner 2013). Almost without exception, experts are calling for more explicit instruction of academic vocabulary, including more focused lists of 'core' academic vocabulary (our current study), as well as lists specific to certain disciplines of education (e.g. history, science, philosophy, political science).

Given the size of the academic vocabulary task—for example, average high school graduates know 75,000 words (Snow and Kim 2007)—pedagogical word lists will continue to be important in academic settings. Such lists are useful in establishing vocabulary learning goals, assessing vocabulary knowledge and growth, analyzing text difficulty and richness, creating and modifying reading materials, designing vocabulary learning tools, determining the vocabulary components of academic curricula, and fulfilling many other crucial academic needs (cf. Nation and Webb 2011). Because so much is currently based on pedagogical word lists, it is crucial that the words in any academic list be truly representative of contemporary academic language, and that they be identified using sound methodological principles—which brings us to our present study involving a new Academic Vocabulary List (AVL).

2. THE NEED FOR A NEW AVL

During the 1970s, pioneering scholars in the area of vocabulary produced several lists of general academic words based on various small corpora of academic materials, primarily consisting of textbooks (Campion and Elley 1971; Praninskas 1972; Lynn 1973; Ghadessy 1979). Because of limited computing power at the time, these lists were compiled by hand. Some were based on basic frequency and range criteria (Campion and Elley 1971; Praninskas 1972), while others were based on student annotations of words they did not understand in their textbooks (Lynn 1973; Ghadessy 1979).

In an attempt to produce a more robust AVL, Xue and Nation (1984) combined all four of the lists indicated above into one *University Word List* (UWL). This list was widely used for >15 years, gaining considerable traction in language education and research by the fact that it was associated with early versions of the popular *Vocabulary Profile* and *Range* computer programs (Nation and Heatley 1994). These user-friendly programs were produced and freely distributed by Nation and his colleagues. However, the need for a more representative academic list was expressed by Coxhead (2000) in a seminal article describing her *Academic Word List* (AWL):

> . . . as an amalgam of the four different studies, it [the UWL] lacked consistent selection principles and had many of the weaknesses of

> the prior work. The corpora on which the studies were based were small and did not contain a wide and balanced range of topics. (p. 214).

Because of its strengths when compared with earlier lists, Coxhead's AWL became the new standard and has served well as a vocabulary workhorse in English language education for over a decade (Coxhead 2011). Recently, the AWL has also received a great deal of attention in primary and secondary education, particularly in the USA (e.g. Hiebert and Lubliner 2008; Baumann and Graves 2010; Nagy and Townsend 2012), where concern continues regarding the widening gap between high and low academic achievers. However, all of this interest in the AWL has also resulted in more careful scrutiny of the methodology behind the list, with several concerns being consistently pointed out in the literature. We will briefly address the two that appear to be most problematic: the use of word families to determine word frequencies, and the relationship of the AWL with the *General Service List* (GSL; West 1953).

2.1 Word families used for initial AWL counts

The AWL was determined by using word families, with a word family being defined as a stem (headword) plus all inflections and transparent derivations containing that stem (Coxhead 2000). For example, the word family *react* contains the following members:

react (headword)
reacted
reacts
reacting

reaction
reactions
reactionaries
reactionary

reactive
reactivate
reactivation
reactor
reactors

The choice to base text coverage on word families has been criticized on several levels. First, members of an extensive word family like *react* may not share the same core meaning (c.f. Nagy and Townsend 2012). Consider, for example, the differences in primary meanings between ***react*** (respond), ***reactionary*** (strongly opposed to social or political change), ***reactivation*** (to make something happen again), and ***reactor*** (a device or apparatus). These meaning differences are accentuated further as members of word families cross over the various academic disciplines (Hyland and Tse 2007).

Many of the meaning problems are caused by the fact that 'word family' does not consider grammatical parts of speech (e.g. nouns, verbs, adjectives, adverbs), as we can see when we analyze a typical AWL word family like *proceed*. We have added the parts of speech for discussion purposes.

> ***proceed*** (verb), ***proceeds*** (verb or noun?), ***procedural*** (adjective), ***procedure*** (noun), ***procedures*** (noun), ***proceeded*** (verb), ***proceeding*** (verb), ***proceedings*** (noun).

Without grammatical identification, the verb ***proceeds*** (meaning *continues*, and pronounced with stress on the second syllable) and the noun ***proceeds*** (meaning *profits*, and pronounced with stress on the first syllable) would be counted as being in the same word family. Many such inaccuracies could be eliminated by counting lemmas (words with a common stem, related by inflection only, and coming from the same part of speech).

Using lemmas would also take care of three additional problems with the *proceed* family above: (i) the noun ***proceedings*** (meaning *records* or *minutes*) would be correctly counted on its own, (ii) the noun ***procedure*** (meaning *technique*) and its inflected plural form, ***procedures***, would be correctly grouped and counted together on their own; and (iii) the adjective ***procedural*** (meaning *technical* or *routine*) would be correctly counted on its own. However, in a word-family approach, all of these word forms (with their variant meanings and grammatical functions) would be counted together as a single word family.

Another major concern with counting word families instead of lemmas to produce pedagogical word lists is that knowledge of derivational word relationships comes much later than knowledge of inflectional word relationships for most school-aged children and second language adults (see Gardner 2007, for review). In other words, 'knowledge of morphologically complex words such as derived nominals [nouns] and derived adjectives is a late linguistic attainment' (Nippold and Sun 2008: 365). Furthermore, numerous studies have shown that the skill of morphological analysis is largely dependent on learners' existing vocabulary knowledge in the first place—a condition that does not favour those most in need of vocabulary help (see Nagy 2007, for review). In short, it is also clear from these learning perspectives that lemmas (inflectional relationships only) should be preferred to word families (inflectional and derivational relationships) in determining pedagogical word lists, especially if those lists are intended to be used by learners at less than advanced English proficiency (cf. Schmitt and Zimmerman 2002).

2.2 Relationship of AWL to GSL

The AWL was built on top of the GSL (West 1953), with the assumption that the GSL contains words of more general high frequency than the AWL (Coxhead 2000; Nation 2001). Several recent articles have questioned this methodology on the grounds that the GSL is an old list (based on a corpus from the early 1900s), and that the AWL actually contains many words in the highest-frequency lists of the *British National Corpus* (BNC; Nation 2004; Hancioğlu *et al*. 2008; Nation 2008; Cobb 2010; Neufeld *et al*. 2011; Schmitt and Schmitt 2012).

In our own analysis, we found the following distribution of AWL families in the top 4,000 lemmas of a recently published frequency dictionary (Davies and Gardner 2010), which is based on *The Corpus of Contemporary American English* (COCA; Davies 2012).

Table 1: AWL word families in the highest frequency bands of COCA

COCA frequency bands	COCA lemma ranks	Number of AWL word families
1	1–1,000	81
2	1,001–2,000	155
3	2,001–3,000	137
4	3,001–4,000	78
	Total	451

Table 1 indicates that 451 of the 570 AWL word families (79%) are represented in the top 4,000 lemmas of COCA, with 236 (Bands 1 and 2) of the 570 (41%) actually being in the top 2,000 lemmas of COCA. It is important to remember that these COCA lemma groupings (inflections only) are not nearly as extensive as word family groupings (inflections plus derivations), thus making the overlap even more noteworthy. These findings with COCA (American English), along with the findings of the BNC studies cited in the first paragraph of this section (British English), provide strong evidence that (i) the AWL is largely a subset of the high-frequency words of English and should therefore not be thought of as an appendage to the GSL, and (ii) the GSL, as a whole, is no longer an accurate reflection of high-frequency English. Regarding the first point, we draw attention to the fact that the AWL produces good coverage of academic materials precisely because it does contain so many high-frequency words. We have no problem with this fact, only with the way that the GSL–AWL relationship has been explained for purposes of instructional vocabulary sequencing and vocabulary-coverage research in academic contexts.

The counter-side of this problem is that there are many high-frequency academic words in the GSL that were not considered in the AWL (Nagy and Townsend 2012; Neufeld *et al.* 2011). For instance, words like *company*, *interest*, *business*, *market*, *account*, *capital*, *exchange*, and *rate* all occur in the GSL and were therefore not considered in the AWL counts, even though such words have major academic meanings. In short, because the GSL words were excluded from the AWL analysis, there is no easy way to separate the high-frequency academic words in the GSL from the high-frequency words that tend to be important in other areas of focus. These include: (i) word families that are common in fiction, but not in academic text (e.g. *bed*, *cup*, *door*, *eye*, *floor*, *hair*, *hang*, *laugh*, *leg*, *morning*, *nice*, *night*, *pretty*, *pull*, *room*, *shake*, *sit*, *smile*, *window*); (ii) word families that occur much more often in magazines than in academic text (e.g. *baby*, *big*, *car*, *cook*, *cup*, *dog*, *fun*, *glass*, *heat*, *hot*, *lot*, *minute*, *pick*, *ride*, *roll*, *shop*, *stick*, *tonight*); and (iii) word families that occur much more often in newspapers than in academic text (*beat*, *big*, *finish*, *game*, *gun*, *hit*, *night*, *park*, *police*, *run*, *sale*, *season*, *shoot*, *stock*, *street*, *throw*, *week*, *win*). We could make similar lists for other non-academic genres, and there are hundreds of such non-academic word families in the GSL. As a preparatory stage for focused academic vocabulary training, sifting through such a list of general English vocabulary (reflecting many genres of English, not just academic) is simply inefficient.

3. IS THERE A 'CORE' ACADEMIC VOCABULARY?

It is important to note that some experts have begun to question whether a core academic list such as the AWL or our proposed list is a viable notion at all. No one seems to dispute the fact that discipline-specific (technical) words are essential to academic understanding, but the value of identifying core academic words that provide useful coverage across a range of different academic disciplines has been questioned on the grounds that such words may change meanings when they cross those disciplines (Hyland and Tse 2007), and that a list of such words may serve no useful purpose in distinguishing between general high-frequency words and academic high-frequency words (Hancioğlu *et al.* 2008; Neufeld *et al.* 2011).

In our view, the first criticism involving meaning-variation could be applied to any high-frequency list of English that is based on forms of words, and on different disciplines of the language (Gardner 2007). In fact, the more general the high-frequency list, the more problems there will be with word-meaning variation because the highest frequency content words of English are the most polysemous (Ravin and Leacock 2000). The question is, until we develop a highly accurate computer program for tagging lexemes in electronic text (word forms and their distinct meanings), should we throw out generalized word lists altogether? From our perspective, the answer to this question is 'no'. Any well-conceived list of high-coverage words brings some order to what otherwise would be vocabulary chaos (Where do we start? What can our learners focus on now, next, etc.?). Until we have an accurate lexeme tagger, we can minimize to some degree the meaning problems by counting lemmas instead of word families, by teaching learners how to deal with multiple meanings, and by providing useful application tools for words on our lists.

Regarding the second issue of no useful distinction between academic and general high-frequency words, we would counter with three points. First, recent research suggests that general academic vocabulary knowledge does make a significant additional contribution to academic achievement (Townsend *et al.* 2012). Secondly, we are not bound in our methodology by any preconceived notion of one list being built on top of another, or that an academic list should somehow follow a general high-frequency list, as was the case with Coxhead's AWL, nor are we concerned with the fact that many core academic words may appear in the highest frequency lists of the BNC, COCA, or any other large corpus of English—in fact, we fully expect it. Finally, it is crucial to identify a statistically viable list of core academic words that can be focused on in academic training and research, rather than losing sight of such words in a new general high-frequency list, or what some have termed 'a face-lifted GSL' (Eldridge 2008: 111).

Major proponents of this concept have recently proposed a new single list for academic settings—the *Billuroğlu–Neufeld List* (*BNL*; Neufeld and Billuroğlu 2005; Hancioğlu *et al.* 2008)—which is accessible on *Compleat Lexical Tutor* (Cobb 2012), and has been adapted on the same site for word-sorting applications. In short, the BNL is an amalgam of several different lists, which we summarize here, based on an article by Hancioğlu *et al.* (2008: 466):

1. The GSL word families
2. The AWL word families
3. The first 2,000 words of the Brown Corpus
4. The first 5,000 words of the BNC (presumably lemmas, but unclear)
5. The 1995 Bauman revision of the GSL
6. The Longman Wordwise commonly used words
7. The Longman Defining Vocabulary

Key to the BNL concept is that the combining and filtering of several established lists should result in a new general list that is less prone to the problems associated with any of the lists individually, and that nearly all of the AWL words are simply 'absorbed' (p. 466) into the new list, essentially eliminating any need to distinguish between a GSL and an AWL.

While we acknowledge the effort to describe the true place of the AWL in terms of its relationship to general high-frequency words, we do not view the amalgamated BNL as the answer for academic settings because it takes us back to the notion of a general list only, in which core academic words are stirred back into the same old polysemous word soup, where 2,000 or 3,000 general high-frequency word families could actually mean something along the lines of 20,000–50,000 lexemes (words with their variant meanings)—numbers that would discourage even the most ambitious learners and teachers in academic settings.

We also believe that the healthy moves toward more specialized AWLs (Hyland and Tse 2007) and the identification of 'general English words that have specific subject meanings' (Eldridge *et al.* 2010: 88) will first require a better understanding of what is common or core before we can determine what is specialized or specific. Again, if the goal is academic English, the statistics can and do point to a narrower list of core academic words that can be focused on by learners, teachers, and researchers.

One final concern with the BNL approach for academic applications is that the only BNL sublist with direct ties to an 'academic' corpus is the AWL. The rest of the sublists were derived from general corpora. In fact, we are puzzled why the BNL developers (and Coxhead as well) did not take advantage of the readily available academic subcorpus of the BNC, consisting of roughly 16 million words (plus many more potential academic sources in the technical magazines of the BNC). Because of the relative non-academic focus of the BNL, the six BNL frequency bands will likely not match the frequency distributions of an actual ‘academic’ corpus, and thus be less informative than is needed for focused academic purposes. Our preliminary analyses support this assertion.

4. KEY CONSIDERATIONS FOR A NEW AVL

Based on our review of the literature and our own research, we believe that a new list of academic core words must consist of the following characteristics:

1. The new list must initially be determined by using lemmas, not word families. Subsequent groupings of the list into families may be warranted for certain instructional and research purposes.
2. The new list must be based on a large and representative corpus of academic English, covering many important academic disciplines.
3. The new list must be statistically derived (using both frequency and dispersion statistics) from a large and balanced corpus consisting of both academic and non-academic materials. The corpus must be large enough and the statistics powerful enough to be able to separate academic core words (those that appear in the vast majority of the various academic disciplines) from general high-frequency words (those that appear with roughly equal and high frequency across all major registers of the larger corpus, including the academic register), as well as from academic technical words (those that appear in a narrow range of academic disciplines).
4. The academic materials in the larger corpus, as well as the non-academic materials to which it will be compared, must represent contemporary English, not dated materials from 20 to 100 years ago. Otherwise, the validity of the new list could be questioned.
5. The new list must be tested against both academic and non-academic corpora, or corpus-derived lists, to determine its validity and reliability as a list of core academic words.

5. METHODOLOGY

In this section, we will first discuss the corpus that was used as the basis of our AVL. We will then discuss how we created the vocabulary list based on the frequency data from the corpus, paying special attention to the different criteria that were used to define ‘core’ academic words.

5.1 Creating the academic corpus

Our academic corpus is both significantly larger and more recent than the corpus that was used for the AWL (Coxhead 2000). The AWL is composed of 3.5 million words of academic texts, with the majority being published in New Zealand. It is nearly evenly divided into four disciplines—arts, commerce, law, and science—with roughly 875,000 words in each discipline. All of the texts are from the early 1960s to the late 1990s.

Our corpus contains >120 million words of academic texts, taken from the 425-million-word COCA (http://corpus.byu.edu/coca). In other words, our corpus is nearly 35 times larger than the AWL corpus. The academic portion of our corpus is also 20% larger than the entire BNC. All texts in our academic corpus were published in the USA, and are representative of written academic materials. Like the corpus used for the AWL, there are no samples of spoken academic language in our corpus, and we acknowledge that some variation in our findings could result from this limitation.

The corpus is composed of the nine disciplines displayed in Table 2. As indicated, 85 million of the 120 million words come from academic journals. Originally, this was the entire academic corpus for our vocabulary list. However, we later decided to 'soften' the journal-heavy corpus by adding ~31.5 million words from academically oriented magazines for all disciplines except Humanities and Education, where there are few topic-specific magazines. Finally, we added roughly 7.5 million words from the finance sections of newspapers to augment the 'Business and Finance' discipline. This was the only discipline that included texts from newspapers, and they were added because of the difficulty of efficiently and accurately retrieving text from formula- and table-laden academic journals dealing with topics such as economics and finance.

As part of COCA, the 120-million-word academic corpus was already tagged for grammatical parts of speech (e.g. nouns, verbs, adjective, adverbs) by the CLAWS 7 tagger from Lancaster University. This allowed us to count lemmas instead of word families, bringing us much closer to an accurate assessment of word forms, functions, and meanings. For instance, we could differentiate and count the 28 different forms of the word family *use* (e.g. *use, uses, using, used, reuse, useless, usability*), distinguishing between such lemmas as the verb *used* in *he used a rake* and the adjective *used* in *they bought a used car*. With the academic corpus established and the lemmatization in place, we were prepared to create a list of core academic words.

5.2 Creating the AVL

The goal was to create a core academic list composed of words like those shown in the middle section of Table 3. To do this, we excluded general high-frequency words like those on the left-hand side of the table, as well as discipline-specific (technical) academic words like those on the right-hand side of the table.

The following four criteria were used to distinguish the academic core:

1. Ratio. To eliminate general high-frequency words from our list (e.g. *way, take, good, never*), we specified that the frequency of the word (lemma) must be at least 50% higher in our academic corpus than in the non-academic portion of COCA (per million words). Our academic corpus is composed of 120 million

Table 2: *COCA academic corpus*

Disciplines	Total size	Journals/ Magazines	Representative titles
Education	8,030,324	J: 8,030,324	Journals: *Education, J Instructional Psychology, Roeper Review, Community College Review;* Magazines: (none)
Humanities	11,111,225	J: 11,111,225	Journals: *Music Educators Journal, African Arts, Style, Art Bulletin, Hispanic Review, Symposium;* Magazines: (none)
History	14,289,007	J: 11,792,026 M: 2,496,981	Journals: *Foreign Affairs, American Studies International, J American Ethnic History;* Magazines: *American Heritage, Military History, History Today*
Social science	16,720,729	J: 15,782,359 M: 938,370	Journals: *Anthropological Quarterly, Geographical Review, Adolescence, Ethnology;* Magazines: *National Geographic, Americas*
Philosophy, religion, psychology	12,463,471	J: 6,659,684 M: 5,803,787	Journals: *Theological Studies, Humanist, Current Psychology, Church History, J Psychology;* Magazines: *Psychology Today, Christian Century, U.S. Catholic*
Law and political science	12,154,568	J: 8,514,782 M: 3,639,786	Journals: *ABA Journal, Perspectives on Political Science, Harvard J of Law & Public Policy, Michigan Law Review;* Magazines: *American Spectator, National Review, New Republic*
Science and technology	22,777,656	J: 13,363,151 M: 9,414,505	Journals: *Bioscience, Environment, Mechanical Engineering, Physics Today, PSA Journal;* Magazines: *Science News, Astronomy, Technology Review*
Medicine and health	9,660,630	J: 5,714,044 M: 3,946,586	Journals: *J Environmental Health, Orthopaedic Nursing, American J Public Health;* Magazines: *Prevention, Men's Health, Total Health*
Business and finance	12,824,831	M: 5,256,801 N: 7,568,030	Journals: (none); Magazines: *Forbes, Money, Fortune, Inc., Changing Times*. Newspapers: 'finance' section.
Total	120,032,441	Academic journals: 84,914,694 Magazines: 31,496,816 (Newspapers: 7,568,030: Business and finance only)	

Table 3: Distinguishing the academic core

	Ratio 1.5	Dispersion .80
High frequency (all genres)	Academic 'core'	Academic technical
noun: *way, part*	noun: *process, analysis*	noun: *assessment, regime*
verb: *take, know*	verb: *indicate, establish*	verb: *democratize, oscillate*
adj: *good, small*	adj: *significant, critical*	adj: *rhetorical, lunar*
adv: *never, very*	adv: *highly, moreover*	adv: *semantically, psychologically*

words, while the other 305 million COCA words are non-academic (e.g. fiction, popular magazines). Therefore, ~28% of COCA is academic (120 million/425 million). Setting the minimum word-selection ratio for academic at 1.50 (50% more) allowed us to find words (lemmas) that occurred at least 42% of the time in academic, which is 50% more than expected (28% 'expected' × 1.5 = 42%).

There is nothing particularly special about the 1.5 Ratio, as there is no commonly accepted value for this measure. We performed extensive experimentation with values as high as 2.0 and as low as 1.2, and simply observed which words entered into and left the academic core list as we adjusted the values. At too high of a figure (e.g. 2.0) we would lose words like *system*, *political*, *create*, *require*, and *rate*, all of which have a ratio between 1.5 and 2.0, and which we would argue are representative words for an academic core. On the other hand, if we set the Ratio too low, then we would pull in too many general high-frequency words (left-hand side of Table 3). For example, the following words have a Ratio between 1.3 and 1.5, and were thus excluded from our academic core: *problem*, *work* (n), *must*, *offer* (v), *different*, *large*, *most*, *also*.

2. Range. The word (lemma) must occur with at least 20% of the expected frequency in at least seven of the nine academic disciplines. For example, based on the size of the Education discipline, the word *ancestor* should occur 238 times, but it only occurs 14 times. Fourteen tokens are only 5.9% of the expected value of 238, and certainly less than the threshold of 20% of expected frequency. Education is therefore not counted as one of the nine disciplines for *ancestor*, and it also has less than the 20% expected frequency in Medicine (18.2% of expected) and Business (5.9% of expected). Because it 'fails' in three disciplines—its range (six) is less than the required (seven)—the word does not appear in our academic core. Other words that have a range of only six are *pollution*, *ideology*, *migrant*, *continent*, *capitalism*, and *audit*. Such words seem too technical and discipline-specific to be included in an academic core. As with the 1.50 value for Ratio, there is no research-recommended value for Range. We experimented with several different Range values (e.g. six disciplines at 30%, or seven disciplines at 10%), and the specified values (seven disciplines at 20%) appeared to give the best results.

3. Dispersion. Words (lemmas) in the core must have a Dispersion of at least 0.80. This Dispersion measure (the Julliand 'd' figure; see Julliand and Chang-Rodriguez 1964) shows how 'evenly' a word is spread across the corpus, and it varies from 0.01 (the word only occurs in an extremely small part of the corpus) to 1.00 (perfectly even dispersion in all parts of the corpus). In a certain sense, Dispersion is superior to the Range measure. For example, assume that a word occurs at the 20% expected level in seven of nine disciplines (thus fulfilling Range), but that in two of the seven disciplines (e.g. Science and Medicine) it is much more frequent than in the other five. In this case, Dispersion would likely be below 0.80 and the word would therefore be eliminated from the list. As with the other two measures, there is no research-recommended score for Dispersion, and repeated tests on the data revealed that 0.80 did well in eliminating words like *taxonomy*, *sect*, *microcosm*, *episodic*, *plenary*, *intoxication*, *restorative*, *deterministic*, and *filial* (all have a Dispersion <0.80, and seem to be fairly technical and discipline-specific), while keeping core academic words like *detect*, *relational*, *coercion*, *situational*, *coordinate*, and *simulated*—all of which have a slightly higher Dispersion score between 0.80 and 0.84.

4. Discipline Measure. The final measure was (like #2 and #3) designed to exclude discipline-specific and technical words. It states that the word cannot occur more than three times the expected frequency (per million words) in any of the nine disciplines. For example, *student* occurs in Education about 6.8 times the expected frequency (taking into account the size of the Education discipline); because this is above 3.0, the word was excluded from the academic core. Examples of other words that exceed 3.0 for a given discipline are *teacher*, *education*, *behaviour* (Education), *native*, *Indian*, *alliance* (History), *participant*, *physical*, *sexual*, *gender* (Social Science), *text*, *object*, *reader*, *essay* (Humanities), *regulation*, *provision*, *impose* (Law and Political Science), *risk*, *treatment*, *exercise*, *stress*, *exposure* (Medicine and Health), *moral*, *spiritual*, *self*, *ministry* (Philosophy, Religion, and Psychology), and *scientist*, *software*, *laboratory*, *cluster* (Science and Technology). Again, such words—while certainly academic in nature—seem too specific for a 'core' list. This final measure works well in eliminating them.

In summary, Criterion #1 (Ratio) helps to exclude general high-frequency words from an academic 'core' (the left-hand side of Table 3), while Criteria #2–4 (Range, Dispersion, Discipline Measure) together help to exclude technical words and words that occur mainly in one or two disciplines (the right-hand side of Table 3). The application of these four criteria resulted in our new AVL. For reasons of space in this article, we are not able to provide the entire AVL—words #1–3,000. However, the complete list can be freely accessed at www.academicwords.info. (On this same site, we also offer a tiered list of the highest frequency words of the 120 million-word academic corpus, with AVL words highlighted within the list.) Table 4 provides a sample of the AVL—words #1–500.

Table 4: Top 500 words (lemmas) in the AVL

1. study.n
2. group.n
3. system.n
4. social.j
5. provide.v
6. however.r
7. research.n
8. level.n
9. result.n
10. include.v
11. important.j
12. process.n
13. use.n
14. development.n
15. data.n
16. information.n
17. effect.n
18. change.n
19. table.n
20. policy.n
21. university.n
22. model.n
23. experience.n
24. activity.n
25. human.j
26. history.n
27. develop.v
28. suggest.v
29. economic.j
30. low.j
31. relationship.n
32. both.r
33. value.n
34. require.v
35. role.n
36. difference.n
37. analysis.n
38. practice.n
39. society.n
40. thus.r
41. control.n
42. form.n
43. report.v
44. rate.n
45. significant.j
46. figure.n
47. factor.n
48. interest.n
49. culture.n
50. need.n
51. base.v
52. population.n
53. international.j
54. technology.n
55. individual.n
56. type.n
57. describe.v
58. indicate.v
59. image.n
60. subject.n
61. science.n
62. material.n
63. produce.v
64. condition.n
65. identify.v
66. knowledge.n
67. support.n
68. performance.n
69. project.n
70. response.n
71. approach.n
72. support.v
73. period.n
74. organization.n
75. increase.v
76. environmental.j
77. source.n
78. nature.n
79. cultural.j
80. resource.n
81. century.n
82. strategy.n
83. theory.n
84. product.n
85. method.n
86. goal.n
87. likely.j
88. note.v
89. represent.v
90. general.j
91. article.n
92. similar.j
93. environment.n
94. language.n
95. determine.v
96. structure.n
97. section.n
98. common.j
99. occur.v
100. current.j
101. available.j
102. present.v
103. term.n
104. reduce.v
105. measure.n
106. involve.v
107. movement.n
108. specific.j
109. focus.v
110. region.n
111. relate.v
112. individual.j
113. quality.n
114. establish.v
115. author.n
116. seek.v
117. compare.v
118. growth.n
119. natural.j
120. various.j
121. standard.n
122. example.n
123. management.n

124. scale.n
125. argue.v
126. degree.n
127. design.n
128. concern.n
129. state.v
130. therefore.r
131. examine.v
132. pattern.n
133. researcher.n
134. task.n
135. traditional.j
136. finding.n
137. positive.j
138. central.j
139. act.n
140. impact.n
141. reflect.v
142. recognize.v
143. context.n
144. relation.n
145. maintain.v
146. African.j
147. concept.n
148. discussion.n
149. associate.v
150. design.v
151. particularly.r
152. purpose.n
153. address.v
154. define.v
155. particular.j
156. benefit.n
157. survey.n
158. effective.j
159. apply.v
160. contain.v
161. understanding.n
162. production.n
163. form.v
164. association.n
165. reveal.v
166. range.n
167. affect.v
168. attitude.n
169. status.n
170. necessary.j
171. function.n
172. indeed.r
173. present.j
174. global.j
175. conflict.n
176. achieve.v
177. conduct.v
178. critical.j
179. perform.v
180. discuss.v
181. exist.v
182. improve.v
183. observe.v
184. demonstrate.v
185. unit.n
186. total.j
187. modern.j
188. literature.n
189. result.v
190. experience.v
191. principle.n
192. element.n
193. challenge.n
194. control.v
195. historical.j
196. aspect.n
197. perspective.n
198. basic.j
199. measure.v
200. tradition.n
201. belief.n
202. western.j
203. procedure.n
204. test.v
205. category.n
206. tend.v
207. technique.n
208. outcome.n
209. significantly.r
210. generally.r
211. future.j
212. mean.n
213. importance.n
214. application.n
215. feature.n
216. influence.n
217. basis.n
218. interaction.n
219. refer.v
220. communication.n
221. negative.j
222. primary.j
223. characteristic.n
224. European.j
225. lack.n
226. obtain.v
227. potential.j
228. variety.n
229. component.n
230. following.j
231. access.n
232. contribute.v
233. assume.v
234. express.v
235. tool.n
236. promote.v
237. participate.v
238. labor.n
239. engage.v
240. review.n
241. additional.j
242. highly.r
243. appropriate.j
244. publish.v
245. encourage.v
246. successful.j
247. assess.v
248. view.v
249. client.n

250. instrument.n
251. relatively.r
252. meaning.n
253. limit.v
254. increase.n
255. directly.r
256. previous.j
257. demand.n
258. vision.n
259. female.j
260. attempt.n
261. influence.v
262. independent.j
263. solution.n
264. direct.j
265. conclusion.n
266. presence.n
267. scientific.j
268. ethnic.j
269. complex.j
270. active.j
271. male.n
272. claim.n
273. participation.n
274. focus.n
275. contrast.n
276. failure.n
277. internal.j
278. journal.n
279. multiple.j
280. facility.n
281. user.n
282. emerge.v
283. protection.n
284. extent.n
285. male.j
286. mental.j
287. explore.v
288. consequence.n
289. generate.v
290. content.n
291. device.n
292. requirement.n
293. broad.j
294. observation.n
295. visual.j
296. difficulty.n
297. regional.j
298. perceive.v
299. ie.r
300. urban.j
301. female.n
302. capacity.n
303. increased.j
304. ensure.v
305. select.v
306. moreover.r
307. emphasize.v
308. institute.n
309. extend.v
310. connection.n
311. sector.n
312. commitment.n
313. interpretation.n
314. evaluate.v
315. conclude.v
316. notion.n
317. increasingly.r
318. domestic.j
319. consist.v
320. reference.n
321. initial.j
322. adopt.v
323. comparison.n
324. depend.v
325. attempt.v
326. standard.j
327. predict.v
328. employ.v
329. definition.n
330. essential.j
331. contact.n
332. frequently.r
333. colleague.n
334. actual.j
335. account.v
336. dimension.n
337. theme.n
338. largely.r
339. link.v
340. desire.n
341. overall.j
342. useful.j
343. consistent.j
344. distribution.n
345. minority.n
346. analyze.v
347. range.v
348. psychological.j
349. unique.j
350. experiment.n
351. trend.n
352. exchange.n
353. percentage.n
354. objective.n
355. implication.n
356. contribution.n
357. enable.v
358. organize.v
359. specifically.r
360. currently.r
361. emotional.j
362. locate.v
363. primarily.r
364. scholar.n
365. enhance.v
366. improvement.n
367. flow.n
368. estimate.v
369. phase.n
370. rural.j
371. typically.r
372. above.r
373. long-term.j
374. core.n
375. volume.n

376. approximately.r
377. limited.j
378. propose.v
379. framework.n
380. existing.j
381. creation.n
382. code.n
383. emphasis.n
384. industrial.j
385. external.j
386. waste.n
387. potential.n
388. climate.n
389. explanation.n
390. technical.j
391. mechanism.n
392. description.n
393. vary.v
394. reduction.n
395. discipline.n
396. construct.v
397. equal.j
398. origin.n
399. rely.v
400. fundamental.j
401. transition.n
402. assumption.n
403. German.j
404. existence.n
405. formal.j
406. manner.n
407. assistance.n
408. combination.n
409. increasing.j
410. hypothesis.n
411. phenomenon.n
412. planning.n
413. error.n
414. household.n
415. cite.v
416. lack.v
417. judgment.n
418. constitute.v
419. relevant.j
420. typical.j
421. selection.n
422. incorporate.v
423. illustrate.v
424. cycle.n
425. depression.n
426. consideration.n
427. previously.r
428. arise.v
429. developing.j
430. separate.j
431. recognition.n
432. mode.n
433. similarly.r
434. resistance.n
435. furthermore.r
436. diversity.n
437. practical.j
438. anxiety.n
439. acquire.v
440. characterize.v
441. differ.v
442. review.v
443. interpret.v
444. creative.j
445. limitation.n
446. resolution.n
447. implementation.n
448. numerous.j
449. significance.n
450. revolution.n
451. philosophy.n
452. display.v
453. professional.n
454. publication.n
455. variation.n
456. derive.v
457. alternative.n
458. widely.r
459. permit.v
460. alternative.j
461. merely.r
462. initiative.n
463. employment.n
464. regard.v
465. estimate.n
466. effectively.r
467. cooperation.n
468. transform.v
469. absence.n
470. imply.v
471. comprehensive.j
472. observer.n
473. nevertheless.r
474. testing.n
475. link.n
476. evolution.n
477. intellectual.j
478. signal.n
479. passage.n
480. facilitate.v
481. discovery.n
482. biological.j
483. introduction.n
484. boundary.n
485. substantial.j
486. ratio.n
487. strongly.r
488. theoretical.j
489. gain.n
490. general.r
491. settlement.n
492. independence.n
493. yield.v
494. formation.n
495. insight.n
496. territory.n
497. conventional.j
498. inform.v
499. index.n
500. crucial.j

Table 5: Word family example for 'define'

98	Define	37,705	**define** (v) $_{23,125}$ **definition** (n) $_{11,955}$ **redefine** (v) $_{1296}$ defining (j) $_{702}$ **defined** (j) $_{563}$ **redefinition** (n) $_{341}$ **undefined** (j) $_{170}$ *definitional* (j) Edu $_{165}$ **definable** (j) $_{110}$ **predefined** (j) $_{104}$ **redefined** (j) $_{41}$ *indefinable* (j) Hum $_{36}$ undefinable (j) $_{17}$ *redefining* (n) Edu + Hum $_{13}$

Word family rankings and frequencies are based on core academic words (lemmas) only, shown here as bold entries.

5.3 Creation of AVL word families

To make direct comparisons with the AWL and other word-family lists, it was necessary to convert our lemma-based AVL into word families, the top 2,000 of which can also be found at www.academicwords.info. Word families were created for the AVL with the aid of Paul Nation's 20,000+ word families, which were merged into our database. Table 5 contains an example entry for our word family *define*.

We suggest that our format for word families includes a number of useful features for instruction and research not available in standard word-family presentations.

1. The families are ordered by frequency (#1 through ~#2,000), and we provide the actual frequencies for each, based on our 120-million-word academic corpus. For example, the family *define* is #98 in the family list and it occurs 37,705 times in the academic corpus.
2. We group the words by lemma, so that *define*, *defines*, *defined*, and *defining* (for the verb *define*) are not listed separately.
3. We separate by part of speech, which provides insight into meaning and usage. For example, we separate *defining* as an adjective and *defining* as part of the verbal lemma *define*.
4. We indicate which list the lemmas are part of. Users can see whether the lemma is part of the core academic list (e.g. *define* and *definition*), technical/discipline-specific (e.g. *definitional* and *indefinable*), or whether it is a lemma that is not academic per se, but is still included in the word family for potential learning purposes (e.g. *defining* as an adjective). (Note: in the downloadable version, these distinctions are color-coded, but in this printed version, we use bold for core academic and italic for technical words.)
5. The AVL word families indicate the frequency of each lemma, allowing learners and teachers to focus first on the most frequent academic lemmas in a particular family. This level of instructional detail is not possible with the AWL and modern versions of the GSL.
6. For those words that are technical/discipline-specific, we indicate the discipline(s) in which they are most common (e.g. *definitional* in Education, or *indefinable* in Humanities).

Table 6: COCA genre coverage of two lists of 570 randomly selected word families

Genres	Genre size	Words in random list 1		Words in random list 2		Average
	# Words	# Words	Coverage	# Words	Coverage	Coverage
Academic	120,847,709	28,021,249	**23.2%**	18,890,254	**15.6%**	**19.4%**
Newspaper	77,553,000	20,444,342	**21.4%**	12,742,459	**16.4%**	**21.4%**
Fiction	83,369,907	23,358,359	**20.6%**	11,064,617	**13.3%**	**20.6%**

6. CASE STUDIES

To test the viability of the new AVL, we carried out several case studies, which we report here. It is important to note that only the top 570 AVL word families were used in the case studies. This allowed us in some cases to make direct comparisons with the 570 word families of the AWL.

6.1 Coverage across genres

Our first case study was to determine whether the AVL is what we are claiming it to be—a reliable list of *academic* rather than *general* English, thus validating both the list and the methodology used to produce it. To test this, we first created two separate word lists, each with 570 randomly selected COCA word families with high overall frequencies. It is important to note that these randomly generated lists included many general high-frequency words of English and will therefore show higher overall coverage figures than our AVL, which is narrowly focused on academic vocabulary.

Table 6 gives the coverage data of the two random lists for three major genres of COCA: academic, newspaper, and fiction. As we would suspect, the two randomly generated lists do not show any marked differences in coverage between the three genres, and certainly would not be good lists for focused academic English purposes, as learners would likely need to sift through many words that are not particularly useful in meeting academic needs.

Compare these nearly equal coverage distributions with those in Table 7 for our AVL in both COCA and the BNC. Note that 'Academic' in the BNC is a combination of all texts marked [w_acad*] and [w_non_ac*] (=technical magazines) in the BNC metadata. This academic subcorpus consists of selections from books (602), journals (367), and other manuscripts (66).

It is clear that in both corpora we get precisely what we would expect of a truly 'academic' list—much better coverage in academic than in newspaper for our AVL, and even better coverage when compared with fiction. This is to be

Table 7: COCA and BNC genre coverage with the AVL

Genres	COCA			BNC		
	Genre size	# Words AVL	Coverage	Genre size	# Words AVL	Coverage
Academic	120,847,709	16,633,796	**13.8%**	32,828,961	4,507,211	**13.7%**
Newspaper	77,553,000	6,229,359	**8.0%**	10,638,034	740,065	**7.0%**
Fiction	83,369,907	2,862,093	**3.4%**	16,194,885	548,708	**3.4%**

Table 8: Coverage of AVL and AWL in COCA academic and BNC academic

List	COCA academic			BNC academic		
	Genre size	# Words	Coverage	Genre size	# Words	Coverage
AVL (570)	120,847,709	16,633,796	**13.8%**	32,828,961	4,507,211	**13.7%**
AWL (570)	120,847,709	8,601,839	**7.2%**	32,828,961	2,261,469	**6.9%**

expected because fiction is much less like academic than newspapers. It is also interesting to note how similar the AVL coverage is in both COCA and the BNC.

6.2 Comparing AVL to AWL

The next case study was designed to compare AVL and AWL coverage in substantial academic corpora. Again, we only used the top 570 word families of our AVL to make a valid comparison with the 570 word families of the AWL.

The results in Table 8 indicate that the AVL has nearly twice the coverage as the AWL. Perhaps equally important is the fact that this advantage holds for both COCA (an American corpus, from which the AVL was derived) and the completely unrelated BNC (a British corpus, which was not used in the creation of the AVL). The results also indicate that there are fundamental differences in the composition of the two lists.

6.3 Is AVL versus AWL a fair comparison?

Some may argue that our AVL–AWL comparisons are inherently 'unfair' because the AWL was placed on top of the GSL, while our list has no such restrictions and may therefore contain many more words with higher general frequencies. To this, we would respond that (i) as discussed previously, the AWL itself actually contains many high-frequency words that were not accounted for because of the decision to place it on top of the dated GSL, and (ii) our comparisons with the AWL were done solely out of the necessity to justify and establish a new academic list, given the important contributions of the AWL over the past decade. In one sense, the AWL and the AVL actually represent different conceptualizations of what 'core academic' means. If our AVL contains more high-frequency words than the AWL, it is because the objective, statistical data suggested that they are 'academic'. We were not constrained by a GSL or any other list.

We also acknowledge that we have benefitted from advancements in technology, corpus construction, and corpus size because the AWL was produced over a decade ago. In short, we view all of these issues as justification for our efforts in creating a new AVL, and also the reason we are not able to simply combine our list with the AWL, or use our list to adjust the AWL, as some have suggested.

Others may argue that it makes no difference whether the AWL follows the GSL because learners will eventually get most of the 'academic' words one way or the other. We see this as a dubious argument if the learning goal is 'academic', not 'general' English. As previously discussed, such a stance requires learners and teachers to sift through a great deal of non-academic and dated vocabulary in the GSL, and also to neglect important academic words they may need now, not later.

6.4 What to do with the new AVL

From our perspective, the key to using the AVL is to focus on English for academic purposes, not English for general purposes. For instance, many learners in academic settings will already have a basic English vocabulary in place (similar to what the GSL represents), and they may even know some words on the AVL. For such learners, the target may be the unknown words in the expanded AVL (3,000 lemmas, 2,000 word families), together with key discipline-specific vocabulary.

For those academic learners who are true beginners in English, there are many long-standing resources such as the Dolch List (1948) and Fry List (1996) that could be learned as a precursor to working with the AVL. As mentioned previously, we are also making freely available at www.academic-words.info the word frequency listing of the 120-million-word academic subcorpus from which the AVL was derived, with AVL words highlighted in the list. Working with the top tiers of this frequency listing (top 500, top 1,000, etc.) would seem to be the best approach for dealing with the vocabulary of beginning learners in academic settings because the list comes from academic, not general texts. Thus, it would contain the standard high-frequency function words found in almost all lists, plus the highest frequency content words of a substantial academic subcorpus (120 million words), some of which will be AVL words.

Finally, the new AVL will be an integral part of a new online resource found at www.wordandphrase.info/academic. There are two main parts to this site.

The first allows users to enter or select AVL words and obtain important data about that word, including (i) synonyms, (ii) definitions, (iii) relative frequency across nine academic disciplines, (iv) the top collocates of the word, which provide useful insights into meaning, usage, and phrasal possibilities, and (v) up to 200 sample concordance lines.

The second part of the site allows users to input a text of their choosing and see frequency information for AVL words and more technical academic words. Additionally, all words and phrases in the text window are searchable in the dynamic ways described above.

7. CONCLUSION

In this article, we have presented the rationale and methodology behind our new AVL. The entire list is available at www.academicwords.info in two formats (a lemma version and a word-family version) to meet various academic needs. The key here is that lemmas, not word families, were used to make initial counts and analyses, and then word families were formed from the lemmas to support certain academic needs. Like all such lists, there are bound to be flaws and limitations. For instance, we recognize that more needs to be done in the future to identify core multiword academic vocabulary (e.g. Simpson-Vlach and Ellis 2010; Martinez and Schmitt 2012), as well as core spoken academic vocabulary (Simpson *et al*. 2002). Acknowledging such limitations, we nonetheless believe the AVL to be the most current, accurate, and comprehensive list of core academic vocabulary in existence today. We have also integrated the list into a dynamic interface at www.wordandphrase.info/academic that will give anyone with internet access the ability to study these words, and to work with them in self-inputted texts. Our hope is that the AVL will be used for the purposes we intended—to improve the learning, teaching, and research of English academic vocabulary in its many contexts.

REFERENCES

Bauman, J. F. and **M. F. Graves.** 2010. 'What is academic vocabulary?,' *Journal of Adolescent and Adult Literacy* 54: 4–12

Beck, I. L., M. G. McKeown, and **L. Kucan.** 2002. *Bringing Words to Life: Robust Vocabulary Instruction*. The Guilford Press.

Biemiller, A. 1999. *Language and Reading Success*. Brookline Books.

Biemiller, A. 2010. *Words Worth Teaching: Closing the Vocabulary Gap*. McGraw-Hill.

Campion, M. and **W. Elley.** 1971. *An Academic Word List*. Wellington New Zealand Council for Educational Research.

Carter, R. 2012. *Vocabulary: Applied Linguistic Perspectives*. Routledge.

Chall, J. 1996. *Stages of Reading Development*. 2nd edn. Harcourt Brace.

Cobb, T. 2010. 'Learning about language and learners from computer programs,' *Reading in a Foreign Language* 22: 181–200.

Cobb, T. 2012. *The Compleat Lexical Tutor for Data-driven Learning on the Web* (v.6.2). Montreal University of Quebec, available at http://lexutor.ca/.

Corson, D. 1997. 'The learning and use of academic English words,' *Language Learning* 47: 671–718.

Coxhead, A. 2000. 'A new academic word list,' *TESOL Quarterly* 34: 213–38.

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),' available at http://corpus.byu.edu/coca/. Accessed June 2012.

Davies, M. and **D. Gardner.** 2010. *A Frequency Dictionary of Contemporary American English: Word Sketches, Collocates, and Thematic Lists*. Routledge.

Dolch, E. W. 1948. *Problems in Reading*. Garrard Press.

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: 109–13.

Eldridge, J., S. Neufeld., and **N. Hancioğlu.** 2010. 'Towards a lexical framework for CLIL,' *International CLIL Research Journal* 1: 80–95.

Fry, E. 1996. *1,000 Instant Words*. NTC/Contemporary Publishing Company.

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

Gardner, D. 2013. *Exploring Vocabulary: Language in Action*. Routledge.

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

Goldenberg, C. 2008. 'Teaching English language learners: What the research does—and does not—say,' *American Educator,* Summer, 8–44.

Graves, M. F. 2006. *The Vocabulary Book: Learning and Instruction*. International Reading Association.

Hancioğlu, N., S. Neufeld, and **J. Eldridge.** 2008. 'Through the looking glass and into the land of lexico-grammar,' *English for Specific Purposes* 27: 459–79.

Hart, B. and **T. R. Risley.** 1995. *Meaningful Differences in the Everyday Experiences of Young American Children*. P.H. Brookes.

Hart, B. and **T. R. Risley.** 2003. 'The early catastrophe: the 30 million word gap by age 3,' *American Educator* 27: 4–9.

Hiebert, E. H. and **S. Lubliner.** 2008. 'The nature, learning, and instruction of general academic vocabulary,' in A. E. Farstrup and S. J. Samuels (eds): *What Research Has to Say About Vocabulary Instruction*. International Reading Association, pp. 106–29.

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

Ippolito, J., J. L. Steele, and **J. F. Samson.** 2008. 'Introduction: why adolescent literacy matters now,' *Harvard Educational Review* 78: 1–6.

Jacobs, V. A. 2008. 'Adolescent literacy: putting the crisis in context,' *Harvard Educational Review* 78: 7–39.

Julliand, A. and **E. Chang-Rodriguez.** 1964. *Frequency Dictionary of Spanish Words*. The Hague Mouton.

Lesaux, N. K., M. J. Kieffer, S. E. Faller, and **J. G. Kelley.** 2010. 'The effectiveness and ease of implementation of an academic vocabulary intervention for linguistically diverse students in urban middle schools,' *Reading Research Quarterly* 45: 196–228.

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

Martinez, R. and **N. Schmitt.** 2012. 'A phrasal expression list,' *Applied Linguistics* 33: 299–320.

Nagy, W. 2007. 'Metalinguistic awareness and the vocabulary-comprehension connection,' in R. K. Wagner, A. E. Muse, and K. R. Tannenbaum (eds): *Vocabulary Acquisition: Implications for Reading Comprehension*. The Guilford Press, pp. 52–77.

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

Nation, I.S.P. 2001. *Learning Vocabulary in Another Language*. Cambridge University Press.

Nation, I.S.P. 2004. 'A study of the most frequent word families in the British National Corpus,' in P. Bogaards and B. Laufer (eds): *Vocabulary in a Second Language: Selection, Acquisition, and Testing*. John Benjamins, pp. 3–13.

Nation, I.S.P. 2008. *Teaching Vocabulary: Strategies and Techniques*. Heinle, Cengage Learning.

Nation, I. S. P. and **A. Heatley.** 1994. *Vocabulary Profile and Range: Programs for the Analysis of Vocabulary in Texts.* [software].

Nation, I.S.P. and **S. Webb.** 2011. *Researching and Analyzing Vocabulary*. Heinle, Cengage Learning.

Neufeld, S. and **A. Billuroğlu.** 2005. 'In search of the critical lexical mass: How 'general' in the GSL? How 'academic' is the AWL?,' available at http://wwwgoogle.com/search?-sourceid=chrome&ie=UTF-8&q=in+search+of+the+critical+lexical+mass%3A+How+'general'+. Accessed June 2012.

Neufeld, S., N. Hancioğlu, and **J. Eldridge.** 2011. 'Beware the range in RANGE, and the academic in AWL,' *System* 39: 533–8.

Neuman, S. B. (ed.). 2008. *Educating the Other America: Top Experts Tackle Poverty, Literacy, and Achievement in Our Schools*. P.H. Brookes.

Nippold, M. A. and **L. Sun.** 2008. 'Knowledge of morphologically complex words: a developmental study of older children and young adolescents,' *Language, Speech, and Hearing Services in Schools* 39: 365–73.

Praninskas, J. 1972. *American University Word List*. Longman.

Ravin, Y. and **C. Leacock.** 2000. 'Polysemy: an overview,' in Y. Raven and C. Leacock (eds): *Polysemy: Theoretical and Computational Approaches*. Oxford University Press, pp. 1–29.

Schmitt, N. and **D. Schmitt.** 2012. 'A reassessment of frequency and vocabulary size in L2 vocabulary teaching,' *Language Teaching*, available at www.journals.cambridge.org. Accessed 8 February 2012.

Schmitt, N. and **C. B. Zimmerman.** 2002. 'Derivative word forms: what do learners know?,' *TESOL Quarterly* 36: 145–71.

Schmitt, N., X. Jiang, and **W. Grabe.** 2011. 'The percentage of words known in a text and reading comprehension,' *The Modern Language Journal* 95: 26–43.

Simpson, R. C., S. L. Briggs, J. Ovens, and **J. M. Swales.** 2002. The *Michigan Corpus of Academic Spoken English*. The Regents of the University of Michigan.

Simpson-Vlach, R. and **N. Ellis.** 2010. 'An academic formulas list: new methods in phraseology research,' *Applied Linguistics* 31: 487–512.

Snow, C. E. and **Y. Kim.** 2007. 'Large problem spaces: the challenge of vocabulary for English language learners,' in R. Wagner, A. E. Muse, and K. R. Tannenbaum (eds): *Vocabulary Acquisition: Implications for Reading Comprehension*. The Guilford Press, pp. 123–39.

Townsend, D. and **P. Collins.** 2009. 'Academic vocabulary and middle school English learners: An intervention study,' *Reading and Writing* 22: 993–1019.

Townsend, D., A. Filippini, P. Collins, and **G. Biancarosa.** 2012. 'Evidence for the importance of academic word knowledge for the academic achievement of diverse middle school students,' *The Elementary School Journal* 112: 497–518.

Vacca, R. T. and **J. A. L. Vacca.** 1996. *Content Area Reading*. 5th edn. Harper Collins.

West, M. 1953. *A General Service List of English Words*. Longman, Green.

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

Zimmerman, C. B. 2009. *Word Knowledge: A Vocabulary Teacher's Handbook*. Oxford University Press.

NOTES ON CONTRIBUTOR

Dee Gardner is an Associate Professor of Applied Linguistics and TESOL at Brigham Young University in Provo, Utah. He is the author of two books and many articles on various vocabulary topics. His primary interests are in the areas of vocabulary acquisition, the vocabulary-reading connection, vocabulary for specific purposes, and applied corpus linguistics. *Address for correspondence*: Department of Linguistics and English Language, Brigham Young University, 4061 JFSB, Provo, Utah, USA 84602. <*dee_gardner@byu.edu*>

Mark Davies is Professor of (Corpus) Linguistics at Brigham Young University. He has published four books and more than sixty articles on corpus linguistics, word frequency, and language change and (genre-based) variation. He is also the creator of several corpora that are freely-available at http://corpus.byu.edu, which are used by hundreds of thousands of users each month. *Address for correspondence*: Department of Linguistics and English Language, Brigham Young University, 4061 JFSB, Provo, Utah, USA 84602. <*mark_davies@byu.edu*>